ADAPTIVE WITHDRAWAL OF REFUSAL STEERING

Anonymous authors
Paper under double-blind review

ABSTRACT

Refusal steering can improve the safety of large language models (LLMs) by shifting internal activations in a refusal direction. Existing refusal steering methods keep steering active for a fixed number of output tokens or throughout generation, yet the number of steered tokens substantially affects both safety and utility. Withdrawing steering too early can allow harmful generation to resume, while keeping it active longer than necessary can increase over-refusal and degrade performance on benign prompts. These opposing risks motivate adaptive termination of inference-time safety interventions. We introduce adaptive withdrawal, a method that predicts the earliest time to end refusal steering for each request while maintaining response safety comparable to persistent steering. The prediction is based only on the model's internal refusal state and its change over the first two output tokens. It does not require observing the rest of the response, allowing earlier withdrawal with less impact on model performance. Across multiple LLMs, adaptive withdrawal achieves safety comparable to persistent steering, including when sampled generation continues long after withdrawal. Compared with state-of-the-art activation-steering methods, adaptive withdrawal achieves competitive safety and benign performance while adding substantially less inference overhead.

1 INTRODUCTION

Jailbreak attacks can elicit harmful responses from safety-aligned large language models (LLMs) (Wei et al., 2023; Zou et al., 2023b; Mazeika et al., 2024; Kirch et al., 2025; Ma et al., 2025; Wu et al., 2026). Refusal steering addresses this failure by shifting internal activations in a refusal direction during generation (Zou et al., 2023a; Arditi et al., 2024; Rimsky et al., 2024; Wollschläger et al., 2025). Existing refusal-steering methods use the input prompt either to decide whether to steer or to select the steering strength (Lee et al., 2025; Zhao et al., 2025b), while feedback-based methods construct steering vectors across model layers (Nguyen et al., 2026). These approaches can decide whether and how strongly to steer, but they do not explicitly predict when refusal steering can end for each prompt (Lee et al., 2025; Zhao et al., 2025b; Nguyen et al., 2026). Most refusal-steering methods remain active throughout generation (Arditi et al., 2024; Rimsky et al., 2024; Lee et al., 2025), while related safety-aware decoding methods modify token probabilities during a fixed early window (Xu et al., 2024). Such predetermined intervention schedules could lead to either excessive steering, resulting in over-refusal, or insufficient steering to reliably suppress harmful responses.

Prior work shows that safety alignment is often concentrated in the first few output tokens and that unsafe prefixes can reinforce harmful continuation, making early withdrawal particularly risky (Qi et al., 2025; Zhang et al., 2025). On the other hand, applying a refusal direction after it is no longer needed can increase refusals on harmless prompts (Arditi et al., 2024) and alter behaviours beyond the intended target (Goyal & Daumé III, 2026). These effects can reduce performance on benign evaluations, including benchmarks designed to measure over-refusal (Röttger et al., 2024; Cui et al., 2025). These risks underscore the lack of a practical policy for determining when steering is no longer required and ordinary decoding can resume without reducing safety.

In this work, we introduce adaptive withdrawal, a method that predicts the earliest time to end refusal steering for each request while maintaining response safety comparable to persistent steering. By ending refusal steering at the predicted time, adaptive withdrawal reduces over-refusal and capability degradation while lowering computational overhead. Using internal activations from the first two output tokens, adaptive withdrawal extracts two signals, namely *refusal veracity* and *refusal velocity*. Refusal veracity measures how strongly the model's internal state expresses refusal at the second

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.

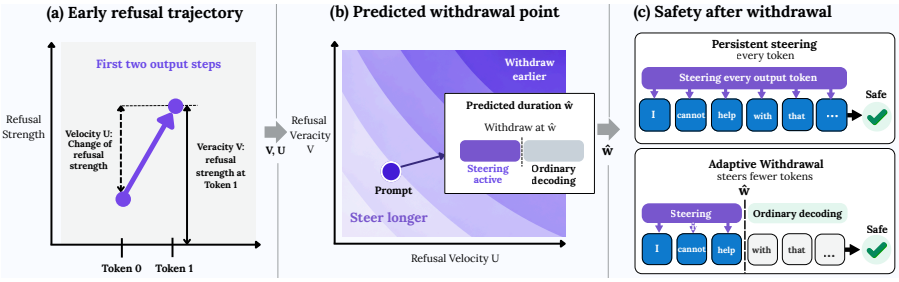


Figure 1: Overview of adaptive withdrawal. (a) Refusal veracity and velocity capture the early refusal trajectory. (b) This trajectory predicts a prompt-specific refusal steering duration. (c) Withdrawing at the predicted time can maintain safety comparable to persistent steering with fewer steered tokens.

output step, while refusal velocity measures how much this signal changes between the first and second steps. This prediction does not require observing the full response or running a safety evaluator during decoding. Figure 1 illustrates the early refusal trajectory, its mapping to a prompt-specific withdrawal time, and the safety comparison after withdrawal.

Extensive experiments across four LLMs and multiple safety evaluators show that adaptive withdrawal achieves safety comparable to persistent steering. The required withdrawal time varies across models, steering configurations, and decoding procedures. It also changes with the safety evaluator applied to held-out responses, making it setting-dependent rather than a universal model property. On adversarial prompts from the Internal Safety Collapse (ISC) benchmark (Wu et al., 2026), adaptive withdrawal maintains competitive safety relative to persistent steering and state-of-the-art activation-steering methods while requiring substantially fewer interventions and adding negligible inference overhead. On benign prompts, persistent steering can increase over-refusal and degrade task performance, whereas adaptive withdrawal substantially reduces steering exposure and limits these losses across multiple benchmarks (Röttger et al., 2024; Cui et al., 2025; Hendrycks et al., 2021). Adaptive withdrawal can substantially improve inference-time safety while maintaining performance on benign tasks. With average LLM response length nearly tripling from 2024 to 2025 (Aubakirova et al., 2026), steering every output token makes model utility pay even when the model gets no safer.

- We show that refusal steering need not remain active throughout generation to maintain safety. The withdrawal time for refusal steering depends on the model, steering configuration, decoding procedure, and safety criterion. Early withdrawal can preserve the safety achieved by persistent steering, even when generation continues long after steering ends.
- We introduce adaptive withdrawal, a method that predicts the earliest time to end refusal steering for each request while maintaining response safety comparable to persistent steering. The prediction uses refusal veracity and refusal velocity from the first two output tokens. It requires neither the full response nor a safety evaluator during decoding.
- We demonstrate that adaptive withdrawal achieves safety comparable to persistent steering across four LLMs while reducing steering exposure, over-refusal, and capability loss. Compared with state-of-the-art activation-steering methods, it maintains competitive safety and benign performance with substantially lower inference overhead.

2 Related Work

Activation Steering and Refusal Representations. Activation steering controls LLM behaviour at inference time by modifying intermediate activations without updating model parameters. Methods such as Inference-Time Intervention and Contrastive Activation Addition derive behavioural directions from selected attention heads or contrastive residual-stream activations and use these directions to modify activations during generation (Li et al., 2023; Rimsky et al., 2024). For safety applications, Arditi et al. (2024) identify a single direction that mediates refusal behaviour. Subsequent studies identify multiple independent refusal directions and higher-dimensional concept cones and show that harmfulness and refusal can be represented separately (Wollschläger et al., 2025; Zhao et al., 2025a). Refusal representations have also enabled automated direction and layer selection (Siu et al., 2025)

and targeted vector ablation to reduce false refusals (Wang et al., 2025a). However, the effects of activation steering vary across inputs and prompt distributions, and steering can alter behaviours beyond its intended target (Tan et al., 2024; Goyal & Daumé III, 2026; Li et al., 2026).

Conditional and Adaptive Activation Steering. Recent work adapts activation steering through input conditioning, adjustments during generation, and layer-dependent vector construction. CAST and FGAS use prompt activations to decide whether steering is needed (Lee et al., 2025; Luo et al., 2026), while AdaSteer and AlphaSteer adapt the steering strength or vector to reduce unwanted effects on benign inputs (Zhao et al., 2025b; Sheng et al., 2026). Other methods vary steering strength during generation (Scalena et al., 2024) or use gradients from learned energy models (Jiang et al., 2026). PID Steering constructs steering vectors using proportional, integral, and derivative terms across model layers (Nguyen et al., 2026). These approaches adapt whether and how to steer, but do not explicitly predict when refusal steering can end for each prompt. Adaptive withdrawal addresses this gap by predicting a prompt-specific withdrawal time.

Temporal Refusal and Safety Recovery. Safety can change throughout autoregressive generation because alignment is often concentrated in the first few output tokens and unsafe prefixes can reinforce later harmful generation (Qi et al., 2025). Recent methods address this temporal dependence by discarding unsafe partial generations, teaching models to transition from harmful output to refusal at different response positions, or reintroducing alignment cues during generation (Zhang et al., 2025; Yuan et al., 2025; Zhang et al., 2026). These methods focus on restoring safety after unsafe generation has begun. Our work instead examines whether early withdrawal of refusal steering can preserve the safety achieved by persistent steering.

Safety-Aware Decoding. Safety-aware decoding improves safety by modifying the output distribution during generation rather than changing internal activations. Existing methods adjust token probabilities during a fixed early window, refine output distributions through self-evaluation, or switch between safety- and utility-oriented distributions using an auxiliary model (Xu et al., 2024; Liu et al., 2024a; Wang et al., 2025b). Other methods, such as RAIN, use self-evaluation and rewinding to revise unsafe partial generations (Li et al., 2024b). Adaptive withdrawal modifies activations and predicts when refusal steering can end. It does not revise generated text or combine output distributions.

3 Adaptive Withdrawal

Adaptive withdrawal predicts, for each prompt, the earliest time refusal steering can end while maintaining safety comparable to persistent steering. Determining this time directly requires generating and evaluating multiple steering durations, which is impractical during inference. Adaptive withdrawal instead predicts it from the model's early refusal trajectory.

3.1 Problem Statement

Our objective is to predict a withdrawal duration for each prompt while maintaining response safety comparable to persistent steering across the evaluated prompt set. We separately validate safety after withdrawal by generating two continuations from the same steered prefix and decoding state, one without further steering and one with steering maintained. Since the continuations are identical up to withdrawal, this comparison isolates whether continued steering is needed to maintain safety.

We first formalise refusal steering. Let $h_{m,d}^{(\ell)}$ denote the residual activation of model m at layer ℓ while predicting the output token at decoding step d. Following Arditi et al. (2024), we construct a refusal direction $v_m^{(\ell)}$ as the unit-normalised difference between the mean activations elicited by harmful and benign source prompts. We define $\sigma_m^{(\ell)}$ as the standard deviation of these source activations projected onto $v_m^{(\ell)}$. The steering strength α scales the refusal direction in units of $\sigma_m^{(\ell)}$.

A steering schedule (s, w) begins at decoding step s and remains active for w consecutive steps. The activation under this schedule is

$$\widetilde{h}_{m,d}^{(\ell)} = h_{m,d}^{(\ell)} + \alpha\sigma_m^{(\ell)} v_m^{(\ell)}, \qquad s \le d < s + w. \tag{1}$$

Following Arditi et al. (2024), persistent steering applies the refusal-direction update at every output step, while unsteered generation applies no update. After step $s + w - 1$, the update is removed, but

the tokens generated under steering remain in the context for subsequent generation. The withdrawal analysis varies both s and w. Adaptive withdrawal begins steering at $s = 0$ and predicts, for each prompt, the shortest duration w that maintains response safety comparable to persistent steering.

3.2 Measuring Safe Withdrawal

We first measure how steering duration affects safety and utility and whether safety persists after withdrawal. These preliminary results reveal model-specific withdrawal durations and motivate adaptive withdrawal.

Steering Start Time and Duration. We vary when refusal steering begins and how many consecutive output tokens receive it. For each model and steering strength, we generate every harmful prompt under each tested schedule, together with unsteered and persistent-steering controls. For a fixed start time, the *restoration width* is the shortest duration that achieves safety comparable to persistent steering and remains comparable at every longer tested duration. A finite restoration width shows that steering need not remain active throughout generation. Its variation across models and settings, illustrated in Figure 2(a), shows why a universal fixed duration is inappropriate.

Safety After Withdrawal. Full-response comparisons alone cannot isolate what happens after steering ends because different schedules can produce different prefixes. We therefore save the steered prefix and decoding state at each tested withdrawal time, then generate one continuation with steering withdrawn and another with steering maintained. The two branches differ only in whether steering continues. We evaluate both the post-withdrawal continuation and the full response, and consider a withdrawal time sufficient only when safety remains comparable to continued steering in both evaluations. Figure 2(b) illustrates this comparison.

Motivation for Prompt-Specific Prediction. The preliminary experiments in Figure 2 test whether refusal steering can end before generation finishes while preserving safety and limiting utility loss. Figure 2(c) reports HarmBench safety across the tested steering start times and durations. Safety generally increases when steering begins earlier and remains active longer, showing that a finite early intervention can match the safety of persistent steering. Figure 2(d) varies how many tokens receive steering on MMLU. MMLU Accuracy generally decreases as this duration grows, demonstrating the capability cost of steering longer than necessary. We demonstrate that the duration required to match persistent-steering safety differs across models, as illustrated in Figure 2(a). Section 4.2 further analyses these preliminary results in detail.

Directly determining the required withdrawal time for a prompt would require generating a separate response for every tested steering duration. Each response would then need to be evaluated for safety and compared with persistent steering. During inference, however, the system must decide when to stop steering within a single ongoing generation, before these alternative responses and their safety labels are available. Repeating the duration sweep would therefore require multiple generations and an online safety evaluator for every request. These limitations motivate adaptive withdrawal, our prompt-specific method for predicting the required withdrawal duration during a single generation.

3.3 Adaptive Withdrawal

We define the prompt-level withdrawal targets and use refusal veracity and refusal velocity to predict the steering duration for each prompt.

Prompt-Level Targets. For each calibration prompt i, we consider schedules with $s = 0$, where steering begins at the first output token and remains active for w tokens. A width is evaluable if the response reaches the withdrawal point and generates at least one subsequent token without steering. We define W_i as the shortest evaluable width at which HarmBench classifies the complete response as safe. Every longer evaluable width must also be classified as safe, preventing an isolated safe response from setting the target.

If no width meets this condition, we treat W_i as right-censored at the longest evaluable width, meaning only that its true value is larger (Singer & Willett, 1993). HarmBench is used to construct W_i and select the threshold for the primary withdrawal policy (Mazeika et al., 2024).

Refusal Veracity and Velocity. At an output-proximal layer ℓ_o, we construct a refusal direction $v_m^{(\ell_o)}$ and scale $\sigma_m^{(\ell_o)}$ using the same harmful-minus-benign procedure as the steering direction. For

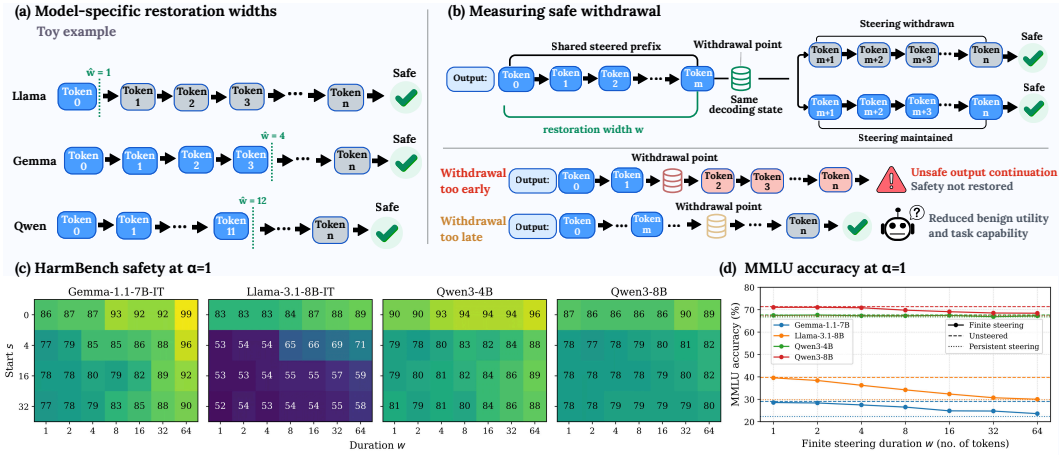


Figure 2: Measuring safe withdrawal and the safety–utility effects of steering duration. (a) Toy examples illustrate that restoration widths differ across models. (b) Paired continuations from the same steered prefix isolate the effect of withdrawal. Ending steering too early can yield an unsafe continuation, whereas continuing it longer than necessary can reduce benign utility and task capability. (c) HarmBench safety at $\alpha = 1$ across refusal-steering start times and durations. Each cell reports the percentage of responses classified as safe. (d) MMLU accuracy (%) at $\alpha = 1$ across steering durations. Dashed and dotted lines show unsteered and persistent-steering performance, respectively.

prompt i, let $\widetilde{h}^{(\ell_o)}_{m,i,d}$ denote the activation produced under steering at decoding step d. We project the activations used to predict the first two output tokens onto the refusal direction:

$$r_{i,d} = \frac{\left\langle \widetilde{h}^{(\ell_o)}_{m,i,d}, v^{(\ell_o)}_m \right\rangle}{\sigma^{(\ell_o)}_m}, \qquad V_i = r_{i,1}, \qquad U_i = |r_{i,1} - r_{i,0}|\,. \tag{2}$$

The activations at $d = 0$ and $d = 1$ are used to predict the first and second output tokens. Refusal veracity V_i measures the strength of the refusal signal at the second step, while refusal velocity U_i measures how much this signal changes between the two steps. Both signals use activations already computed during generation, requiring neither later output tokens or additional model forward pass.

Withdrawal-Time Classifier. We fit a separate discrete-time logistic classifier for each model and steering strength using refusal veracity V_i and refusal velocity U_i. For each tested duration w, the classifier estimates $h_i(w) = \Pr(W_i = w \mid W_i \geq w, V_i, U_i)$, the probability that w is the first duration to reach the prompt-level target, given that no shorter duration has reached it. The classifier is defined by $\operatorname{logit} h_i(w) = a_w + \beta_V z(V_i) + \beta_U z(U_i)$, where $\beta_V, \beta_U \geq 0$. The function $z(\cdot)$ standardises each signal using the calibration-set mean and standard deviation. The duration-specific intercept a_w captures the baseline probability at duration w, while β_V and β_U capture the contributions of refusal veracity and refusal velocity. We constrain the coefficients to be nonnegative, so increasing either signal cannot lower this probability. We fit the classifier on all calibration prompts, including those for which no withdrawal time is observed within the tested durations.

Withdrawal Policy. For each candidate duration w, we use the classifier outputs to compute the cumulative probability $\Pr(W_i \leq w \mid V_i, U_i)$ that the safety target has been reached. After observing the first two output-token steps, adaptive withdrawal selects the earliest candidate duration $w \geq 2$ whose cumulative probability exceeds a threshold chosen using the calibration data. Withdrawal cannot occur until two output tokens have been generated because both prediction signals require two steps. If no candidate duration reaches the threshold, steering remains active throughout generation.

Output steering exposure is the mean number of output steps that receive steering. The withdrawal threshold must satisfy the HarmBench safety criterion relative to persistent steering and reduce steering exposure by at least 25%. Among the thresholds meeting both requirements, we select the one with the lowest exposure. If none qualifies, steering remains active throughout generation. The predictor and threshold are fixed before evaluating the held-out prompts.

4 EXPERIMENTS

We first examine how steering duration affects safety and utility. We then evaluate prompt-specific withdrawal prediction, compare adaptive withdrawal with persistent and state-of-the-art steering, and test its generalisation and robustness to an adaptive attack.

4.1 EXPERIMENTAL SETUP

LLMs. We evaluate Gemma-1.1-7B-IT (Gemma Team et al., 2024), Llama-3.1-8B-Instruct (Grattafiori et al., 2024), Qwen3-4B, and Qwen3-8B (Yang et al., 2025) (see Appendix A for configurations).

Prompts. For each model and steering strength, we pair 30 WildTeaming requests with AIM, AutoDAN, Developer Mode v2, and Evil Confidant, producing 120 calibration prompts (Jiang et al., 2024; Wei et al., 2023; Liu et al., 2024b). The primary $\alpha = 1$ evaluation uses 400 prompts from 100 held-out behaviours under the same four attacks. For out-of-distribution evaluation, we use 96 Internal Safety Collapse (ISC) prompts constructed from StrongREJECT requests and balanced across six harmful-behaviour categories. These prompts require structured responses of at least 120 words (Wu et al., 2026; Souly et al., 2024). We also evaluate 313 direct StrongREJECT requests, 100 GCG prompts, and 83 PAIR prompts (Zou et al., 2023b; Chao et al., 2025). Benign evaluation uses 500 OR-Bench prompts, 250 XSTest prompts, and 1,026 MMLU questions per model (Cui et al., 2025; Röttger et al., 2024; Hendrycks et al., 2021). Appendix A describes each prompt set.

Steering and Predictor Training. We use $\alpha = 1$ as the primary steering strength and $\alpha \in \{2, 3\}$ for the strength analysis. We test start times $\{0, 4, 16, 32\}$ and durations $\{1, 2, 4, 8, 16, 32, 64\}$. For each model and strength, we use the HarmBench evaluator to determine how long steering must remain active for each of the 120 calibration prompts, then fit a predictor using refusal veracity and velocity. We standardise the features and select the withdrawal threshold on calibration data. The predictor minimises an L_2-penalised right-censored negative log-likelihood, with regularisation selected through five-fold cross-validation grouped by harmful behaviour. Appendix A provides further training and decoding details.

Baselines. We compare adaptive withdrawal with unsteered generation, persistent steering (Arditi et al., 2024), and the SOTA activation-steering methods CAST, AdaSteer, and AlphaSteer (Lee et al., 2025; Zhao et al., 2025b; Sheng et al., 2026). We also evaluate a fixed-duration policy selected on the calibration data (Appendix A.3.1) and withdrawal after the first completed refusal cue (Table 15).

Evaluation. HarmBench evaluator is used to construct W_i and select the threshold for the primary withdrawal policy on calibration data (Mazeika et al., 2024). It also evaluates response safety in the primary, ISC, StrongREJECT, GCG, and PAIR experiments. After freezing the policy, we test evaluator transfer using a prompted Qwen2.5-14B-Instruct judge (Yang et al., 2024), Granite Guardian (Padhi et al., 2025), and MD-Judge (Li et al., 2024a). The Qwen judge applies our fixed binary rubric to the harmful behaviour and generated response. We consider withdrawal comparably safe when the one-sided 95% confidence bound rules out a safety reduction greater than 0.05 relative to continued steering. Appendix B provides the evaluator and statistical protocols.

To quantify intervention exposure, we follow Stoehr et al. (2024) and report the mean number of intervention points (MIP) per response. An intervention point is a modified layer–token–site combination. Since each method modifies one activation site per selected layer, the number of intervention points for each response equals the number of selected layers multiplied by the total number of prompt and generated tokens receiving steering. We report the mean across responses.

4.2 SAFETY AND UTILITY ACROSS STEERING DURATIONS

We begin with an exhaustive duration sweep as a preliminary test of whether refusal steering must remain active throughout generation. As shown in Figure 2(c), all four models achieve safety comparable to persistent steering after a finite number of steered tokens when steering begins at the first output token. Qwen3-8B requires 32 tokens, while Gemma, Llama, and Qwen3-4B require 64. Early withdrawal can therefore preserve safety, but the different restoration widths show that the required withdrawal time is model-specific.

Table 1: Safety and benign capability across four LLMs. Values are means over five reruns under greedy decoding, with sample standard deviations in parentheses. We compare persistent steering (Arditi et al., 2024), AdaSteer (Zhao et al., 2025b), AlphaSteer (Sheng et al., 2026), CAST (Lee et al., 2025), and adaptive withdrawal. The in-distribution (ID) held-out set combines AIM, AutoDAN, Developer Mode v2, and Evil Confidant. Safety is also evaluated on ISC, GCG, PAIR, and StrongREJECT. OR-Bench and XSTest report answer rates, while MMLU reports accuracy.

Model	Method	Safety					Benign capability		
		ID held-out	ISC	GCG	PAIR	StrongREJECT	OR-Bench	XSTest	MMLU
Gemma-1.1-7B	Unsteered	75.9 (0.1)	96.5 (0.6)	71.8 (1.8)	91.6 (0.0)	99.2 (0.3)	**96.2 (0.0)**	**80.4 (0.0)**	29.0 (0.0)
	Persistent steering	99.8 (0.0)	99.0 (0.0)	99.0 (1.7)	**96.1 (0.5)**	**99.6 (0.3)**	94.1 (0.1)	61.2 (1.1)	22.0 (0.4)
	AdaSteer	75.5 (0.0)	**100.0 (0.0)**	98.0 (1.9)	80.7 (0.0)	99.2 (0.3)	94.8 (0.0)	80.2 (0.2)	26.2 (0.1)
	AlphaSteer	**100.0 (0.0)**	99.2 (0.5)	99.0 (1.0)	92.8 (0.0)	99.2 (0.5)	94.4 (0.0)	79.0 (0.2)	27.1 (0.4)
	CAST	**100.0 (0.0)**	99.0 (0.0)	**100.0 (0.0)**	**96.1 (0.5)**	99.3 (0.1)	90.6 (0.2)	79.8 (0.2)	23.7 (0.4)
	Adaptive withdrawal	97.0 (0.2)	**100.0 (0.0)**	98.6 (0.5)	**96.1 (0.5)**	**99.6 (0.3)**	96.1 (0.1)	78.8 (0.0)	**29.1 (0.1)**
Llama-3.1-8B	Unsteered	53.3 (0.1)	95.8 (0.0)	84.0 (0.0)	48.2 (0.9)	98.1 (0.0)	**98.4 (0.0)**	93.2 (0.0)	**39.8 (0.0)**
	Persistent steering	90.4 (0.2)	**97.9 (0.0)**	99.0 (0.0)	88.4 (0.7)	98.0 (0.2)	94.3 (0.1)	81.3 (0.2)	29.8 (0.0)
	AdaSteer	92.8 (0.0)	82.3 (0.0)	99.8 (0.4)	86.7 (0.0)	91.4 (0.0)	96.6 (0.0)	**96.8 (0.0)**	39.3 (0.0)
	AlphaSteer	**94.3 (0.3)**	82.3 (0.0)	**100.0 (0.0)**	83.4 (0.5)	97.8 (0.0)	96.3 (0.1)	93.6 (0.0)	38.3 (0.2)
	CAST	90.2 (0.3)	95.8 (0.0)	98.8 (0.4)	**92.0 (0.7)**	97.5 (0.1)	98.2 (0.0)	93.4 (0.9)	27.9 (0.0)
	Adaptive withdrawal	90.2 (0.1)	**97.9 (0.0)**	99.2 (0.4)	88.2 (1.0)	**99.1 (0.1)**	**98.4 (0.0)**	94.0 (0.0)	**39.8 (0.0)**
Qwen3-4B	Unsteered	79.5 (0.0)	89.6 (0.0)	84.0 (0.0)	50.6 (0.0)	98.1 (0.0)	**98.6 (0.0)**	**96.4 (0.0)**	**67.6 (0.0)**
	Persistent steering	95.5 (0.0)	**99.8 (0.5)**	**100.0 (0.0)**	84.3 (0.0)	99.7 (0.0)	98.4 (0.0)	88.6 (0.2)	66.7 (0.1)
	AdaSteer	93.1 (0.3)	70.8 (0.0)	99.0 (0.0)	90.6 (0.5)	**100.0 (0.0)**	82.2 (0.0)	34.4 (0.0)	66.8 (0.1)
	AlphaSteer	98.2 (0.2)	95.6 (0.5)	99.4 (0.5)	84.3 (0.0)	99.0 (0.0)	96.8 (0.0)	92.5 (0.4)	66.6 (0.3)
	CAST	**99.2 (0.1)**	**99.8 (0.5)**	**100.0 (0.0)**	**98.8 (0.0)**	99.7 (0.0)	**98.6 (0.0)**	**96.4 (0.0)**	67.5 (0.2)
	Adaptive withdrawal	95.5 (0.0)	**99.8 (0.5)**	99.2 (0.4)	83.1 (0.0)	99.7 (0.0)	98.4 (0.2)	94.4 (0.0)	**67.6 (0.1)**
Qwen3-8B	Unsteered	78.5 (0.0)	88.5 (0.0)	91.0 (0.0)	44.8 (0.5)	**100.0 (0.0)**	**99.2 (0.0)**	94.8 (0.0)	**71.3 (0.0)**
	Persistent steering	93.6 (0.1)	97.9 (0.0)	**100.0 (0.0)**	91.6 (0.0)	99.9 (0.1)	98.8 (0.0)	94.6 (0.5)	67.2 (0.0)
	AdaSteer	97.8 (0.0)	91.7 (0.0)	99.8 (0.4)	97.3 (0.5)	99.0 (0.0)	90.4 (0.0)	70.7 (1.1)	52.1 (0.3)
	AlphaSteer	98.5 (0.4)	94.6 (0.5)	**100.0 (0.0)**	92.8 (0.0)	99.6 (0.1)	98.4 (0.1)	94.0 (0.0)	69.3 (0.0)
	CAST	**99.0 (0.2)**	**100.0 (0.0)**	**100.0 (0.0)**	**98.3 (0.7)**	**100.0 (0.0)**	97.6 (0.0)	**96.0 (0.0)**	71.2 (0.0)
	Adaptive withdrawal	93.5 (0.1)	93.8 (0.0)	**100.0 (0.0)**	90.4 (0.0)	99.7 (0.0)	**99.2 (0.0)**	95.2 (0.0)	**71.3 (0.1)**

We next examine the corresponding effect of steering duration on model capability. Figure 2(d) shows that MMLU accuracy generally decreases as more analysis tokens receive steering, with persistent steering producing the lowest accuracy for all four models. Continuing steering beyond the required duration can therefore reduce capability without providing additional safety. We also conduct extensive experiments with different steering strengths and safety evaluators under greedy and sampled decoding, finding that the required duration changes across these settings. Appendices C and D report the complete results. These preliminary results establish the feasibility of early withdrawal and motivate selecting the withdrawal time for each prompt.

4.3 Prompt-Specific Withdrawal Prediction

Adaptive withdrawal predicts a prompt-specific steering duration from refusal veracity and refusal velocity. We first evaluate prediction performance across four LLMs and then compare adaptive withdrawal with persistent steering and SOTA activation-steering methods.

Prediction Performance. We evaluate prediction using the concordance index (C-index). Given two prompts requiring different steering durations, the C-index measures how often the predictor correctly identifies which prompt requires longer steering.

Across Gemma, Llama, and Qwen3-8B, the two-token predictor achieves C-indices of 0.919–0.947, correctly ranking approximately 92–95% of prompt pairs. Qwen3-4B is more challenging, with a C-index of 0.669. For Qwen3-8B, adding refusal velocity to veracity raises the C-index from 0.800 to 0.920, showing that early changes in the refusal state provide information beyond its current strength.

To test whether waiting for more evidence improves prediction, we delay the decision until four or eight tokens have been generated. Neither longer observation window improves prompt ranking, and the two-token predictor achieves the highest C-index for every model. Waiting longer also prevents steering from being withdrawn as early. We therefore use the two-token predictor throughout our

Table 2: Mean intervention points and runtime overhead across four LLMs. Following Stoehr et al. (2024), an intervention point is a modified layer–token–site combination. Runtime overhead is the percentage increase relative to unsteered generation.

Model	Method	Mean intervention points								Efficiency
		ID held-out	ISC	GCG	PAIR	StrongREJECT	OR-Bench	XSTest	MMLU	Runtime overhead
Gemma-1.1-7B	Persistent steering	46.2	30.9	31.2	37.5	34.4	120.7	63.4	99.7	0.103
	AdaSteer	13,179.1	26,674.7	2,210.3	2,201.5	1,094.7	856.6	539.3	18,585.8	0.623
	AlphaSteer	3,294.8	6,668.7	552.6	550.4	273.7	214.2	134.8	4,646.4	0.633
	CAST	4,576.0	8,862.7	991.9	917.8	460.6	1,099.7	**7.5**	6,636.1	5.230
	Adaptive withdrawal	**17.3**	**2.0**	**2.8**	**4.5**	**2.0**	**24.5**	14.3	**54.6**	**0.060**
Llama-3.1-8B	Persistent steering	54.5	502.6	40.5	58.7	37.2	124.3	108.4	101.0	0.101
	AdaSteer	16,056.8	27,308.0	3,150.7	3,351.5	2,070.6	1,813.8	1,449.0	21,493.1	2.678
	AlphaSteer	5,017.8	8,533.8	984.6	1,047.3	647.1	566.8	452.8	6,716.6	0.923
	CAST	5,394.8	**0.0**	322.6	1,791.3	1,496.6	1,497.5	708.0	4,992.2	6.971
	Adaptive withdrawal	**33.5**	502.6	**2.8**	**35.1**	**2.1**	**94.3**	**84.8**	**95.9**	**0.038**
Qwen3-4B	Persistent steering	94.7	**288.9**	71.4	118.9	106.4	127.2	119.9	101.0	0.193
	AdaSteer	17,191.3	30,618.4	2,800.3	2,927.4	1,505.6	1,212.8	803.4	23,748.8	1.115
	AlphaSteer	8,590.1	15,309.2	1,402.4	1,471.2	752.8	606.4	401.7	11,874.4	1.152
	CAST	5,446.9	12,262.6	1,835.1	275.9	**9.3**	**0.0**	**0.0**	**0.0**	4.827
	Adaptive withdrawal	**26.1**	**288.9**	**13.4**	**98.2**	22.6	117.9	112.8	95.9	**0.106**
Qwen3-8B	Persistent steering	77.6	506.7	76.9	122.8	122.2	127.8	123.5	101.0	0.149
	AdaSteer	17,170.4	30,618.4	2,804.8	2,942.5	1,505.6	1,212.8	803.4	23,748.8	1.399
	AlphaSteer	8,595.6	15,309.2	1,415.9	1,467.4	752.8	606.4	401.7	11,874.4	1.523
	CAST	4,825.7	12,262.6	1,827.8	1,822.1	1,285.7	1,287.1	**22.2**	6,801.2	4.479
	Adaptive withdrawal	**23.3**	**2.0**	**2.0**	**12.6**	**2.0**	**23.6**	38.6	**3.5**	**0.054**

experiments. Appendix E.2.2 reports the observation-length comparison, while Appendix E provides the complete feature ablations and additional withdrawal baselines.

Comparison with Baselines. Tables 1 and 2 compare adaptive withdrawal with persistent steering and the SOTA methods. On the primary ID evaluation, adaptive withdrawal retains 97.2–100.0% of persistent-steering safety while reducing intervention exposure by 38.5–72.5%. Across all five safety benchmarks, its safety averaged across the four models retains at least 98.9% of persistent steering.

The out-of-distribution evaluations show a similar pattern. Adaptive withdrawal exceeds the average safety of AdaSteer and AlphaSteer on ISC, PAIR, and StrongREJECT, while remaining close to persistent steering. CAST achieves the highest average PAIR safety but requires substantially broader intervention.

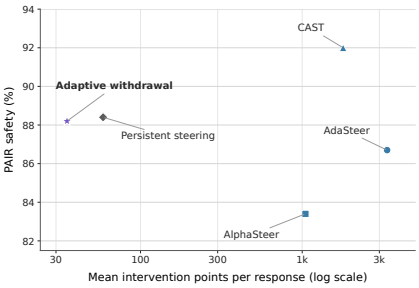


Figure 3: Safety and intervention exposure for Llama-3.1-8B on PAIR.

Figure 3 makes this trade-off concrete. On Llama-3.1-8B, adaptive withdrawal achieves 88.2% PAIR safety using 35.1 intervention points, compared with 88.4% using 58.7 under persistent steering. It also achieves higher safety than AdaSteer and AlphaSteer while using 99.0% and 96.6% fewer intervention points, respectively. CAST reaches 92.0% safety but uses 51 times as many intervention points. Appendix F reports the corresponding trade-offs for every model and evaluation dataset.

Across the complete evaluation, adaptive withdrawal never uses more intervention points than persistent steering and consistently uses fewer than AdaSteer and AlphaSteer. CAST records zero intervention points on the Qwen3-4B benign benchmarks because its conditional gate never activates, so no activation steering is applied. Despite this, CAST still incurs the cost of evaluating its gate. Adaptive withdrawal has the lowest runtime overhead for every model, remaining below 0.11%, compared with 0.62–6.97% for the SOTA methods.

Adaptive withdrawal also preserves benign capability. It matches or improves persistent steering across OR-Bench, XSTest, and MMLU for every model. These results show that adaptive withdrawal preserves the safety benefits of refusal steering while substantially reducing intervention exposure, runtime cost, and capability degradation.

4.4 Generalisation and Robustness

We evaluate adaptive withdrawal across unseen evaluators, models, and steering strengths, followed by an adaptive attack on its withdrawal rule.

Transfer Across Evaluators. To test whether the results transfer beyond HarmBench evaluator, we evaluate responses generated by adaptive withdrawal and persistent steering using three additional safety evaluators. Under the prompted Qwen2.5 judge, adaptive withdrawal retains at least 99.0% of persistent-steering safety and meets the safety criterion for all four LLMs (Yang et al., 2024). Granite Guardian finds comparable safety for three of the four LLMs, with adaptive withdrawal achieving 95.75% safety on Gemma (Padhi et al., 2025). MD-Judge provides more conservative estimates, with adaptive withdrawal retaining 89.7–95.9% of persistent-steering safety but no LLM meets the safety criterion (Li et al., 2024a). These results show that adaptive withdrawal can generalise beyond HarmBench evaluator while maintaining comparable safety under independently defined criteria. Appendix G.1 reports the complete results.

Transfer Across Models and Steering Strengths. Predictors trained on other models without target-model safety labels achieve C-indices of 0.880, 0.946, 0.619, and 0.870 for Gemma, Llama, Qwen3-4B, and Qwen3-8B, respectively. The early trajectory therefore transfers across models, though less reliably to Qwen3-4B. Transferring the predictor and threshold is harder. The policy meets the safety criterion for Llama and Qwen3-8B, but not Gemma or Qwen3-4B. On Qwen3-4B, it still retains 94.0% of persistent-steering safety while reducing exposure by 87.9%. This shows that the early trajectory generalises better than the model-specific withdrawal boundary.

Transfer to stronger steering is more reliable. Applying the frozen $\alpha = 1$ policies at $\alpha \in \{2, 3\}$ meets the safety criterion for Gemma, Llama, and Qwen3-4B, retaining 97.7–100.0% of persistent-steering safety while reducing exposure by 96.1–98.4%. Qwen3-8B reduces exposure by 90.2–94.0% but misses the criterion. Thus, a policy learned at the weakest tested strength can remain effective under stronger steering without recalibration. Appendix G.2 reports the full results.

Adaptive Attack. We test whether an attacker can trigger early withdrawal by making the model refuse briefly before providing harmful content. We construct ten manually written refusal-then-comply prompt variants, each asking the model to begin with a refusal or warning before answering the harmful request. We evaluate 80 attacks per model. For each request, the attacker uses the first two output-token activations to select the variant predicted to trigger the earliest withdrawal. Across the 320 attacks, 318 trigger withdrawal, but only 14 make the withdrawn continuation unsafe when continued steering remains safe. Hence, triggering withdrawal is usually insufficient to bypass refusal. Appendix G.4 provides the complete protocol, results, and safeguard analysis.

5 Discussion

Adaptive withdrawal could be integrated into inference frameworks such as vLLM (Kwon et al., 2023) or SGLang (Zheng et al., 2024) by adding its steering operation and predictor to decoding. For each request, the engine steers the first two output tokens, predicts the withdrawal step from their activations, and stops steering at that step. During continuous batching, a per-request mask lets requests stop steering independently. The predictor uses existing decoding activations and requires no extra forward pass, safety evaluator, or KV-cache recomputation.

6 Conclusion

We introduced adaptive withdrawal, a prompt-specific method for predicting when refusal steering can end while maintaining safety close to persistent steering. It uses refusal veracity and refusal velocity from the first two output steps without observing the full response or invoking a safety evaluator during decoding. Across four LLMs, adaptive withdrawal substantially reduces intervention exposure and limits losses in benign performance with negligible runtime overhead. Our transfer experiments show that the early refusal trajectory generalises more reliably than the complete withdrawal policy, whose decision boundary remains dependent on the model and steering configuration. Future work can investigate whether early internal trajectories provide reliable stopping signals for a broader range of inference-time safety interventions.

AI USE STATEMENT

In this work, we used generative AI tools for [implement methods,support qualitative and thematic data analysis, interpret results, clean and reformat dataset]. We have not used generative AI tools for [help develop theoretical models or conceptual frameworks, propose or refine hypotheses, design or provide feedback on research methodology or experiments], and [Generate synthetic data sets, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, assist with translation] are not applicable to this work. Additionally, we used generative AI tools for [summarize or analyse existing literature, sourcing/searching for information, edit a research paper to improve readability, identify relevant literature]. We have reviewed all AI-assisted work. [AI-generated code was inspected, executed, and tested, reported results were checked against saved experiment outputs, and literature suggestions were verified against their original sources. All AI-assisted text, claims, tables, and figures were manually reviewed and revised by the authors]. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work uses public benchmark prompts and open-weight language models and involves no human participants or personal data. Although adaptive withdrawal is designed to strengthen and safely withdraw refusal interventions, analyses of refusal mechanisms may have dual-use implications. We therefore report aggregate safety results without reproducing actionable harmful responses. Any deployment should include model-specific safety evaluation and appropriate controls against misuse.

REPRODUCIBILITY STATEMENT

Section 3 specifies the adaptive withdrawal methodology, while Section 4.1 describes the evaluation protocol. Appendix A provides model checkpoints, layer choices, prompt splits, calibration procedures, and statistical criteria, while Appendix B documents the evaluator settings. The remaining appendices report complete duration profiles, same-prefix tests, prediction and transfer results, and adaptive-attack evaluations. We provide an anonymised supplementary notebook implementing refusal-coordinate construction, refusal veracity and velocity, and inference-time withdrawal selection.

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 *NeurIPS*, volume 37, pp. 136037–136083, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/f545448535dfde4f9786555403ab7c49-Abstract-Conference.html.

Malika Aubakirova, Alex Atallah, Chris Clark, Justin Summerville, and Anjney Midha. State of ai: An empirical 100 trillion token study with openrouter. *arXiv preprint arXiv:2601.10088*, 2026. doi: 10.48550/arXiv.2601.10088. URL https://arxiv.org/abs/2601.10088.

Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J Pappas, and Eric Wong. Jailbreaking black box large language models in twenty queries. In *SaTML*, pp. 23–42. IEEE, 2025. URL https://ieeexplore.ieee.org/abstract/document/10992337.

Justin Cui, Wei-Lin Chiang, Ion Stoica, and Cho-Jui Hsieh. OR-bench: An over-refusal benchmark for large language models. In *ICML*, volume 267, pp. 11515–11542. PMLR, 2025. URL https://proceedings.mlr.press/v267/cui25a.html.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024. URL https://arxiv.org/abs/2403.08295.

Navita Goyal and Hal Daumé III. Steering safely or off a cliff? rethinking specificity and robustness in inference-time interventions. In *EACL*, pp. 5723–5738, 2026. URL https://aclanthology.org/2026.eacl-long.268/.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. URL https://arxiv.org/abs/2407.21783.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *ICLR*, 2021. URL https://arxiv.org/abs/2009.03300.

Eric Hanchen Jiang, Weixuan Ou, Run Liu, Shengyuan Pang, Guancheng Wan, Ranjie Duan, Wei Dong, Kai-Wei Chang, XiaoFeng Wang, Ying Nian Wu, et al. Mitigating over-refusal in aligned large language models via inference-time activation energy. In *ACL*, pp. 37930–37950, 2026. URL https://aclanthology.org/2026.acl-long.1759/.

Liwei Jiang, Kavel Rao, Seungju Han, Allyson Ettinger, Faeze Brahman, Sachin Kumar, Niloofar Mireshghallah, Ximing Lu, Maarten Sap, Yejin Choi, et al. Wildteaming at scale: From in-the-wild jailbreaks to (adversarially) safer language models. In *NeurIPS*, volume 37, pp. 47094–47165, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/54024fca0cef9911be36319e622cde38-Abstract-Conference.html.

Nathalie Maria Kirch, Constantin Niko Weisser, Severin Field, Helen Yannakoudakis, and Stephen Casper. What features in prompts jailbreak llms? investigating the mechanisms behind attacks. In *Proceedings of the 8th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 480–520, 2025. URL https://aclanthology.org/2025.blackboxnlp-1.28/.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *SOSP*, pp. 611–626, 2023. URL https://dl.acm.org/doi/abs/10.1145/3600006.3613165.

Bruce W Lee, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Erik Miehling, Pierre Dognin, Manish Nagireddy, and Amit Dhurandhar. Programming refusal with conditional activation steering. In *ICLR*, volume 2025, pp. 90960–90985, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/e2dd53601de57c773343a7cdf09fae1c-Abstract-Conference.html.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *NeurIPS*, volume 36, pp. 41451–41530, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/81b8390039b7302c909cb769f8b6cd93-Abstract-Conference.html.

Lijun Li, Bowen Dong, Ruohui Wang, Xuhao Hu, Wangmeng Zuo, Dahua Lin, Yu Qiao, and Jing Shao. Salad-bench: A hierarchical and comprehensive safety benchmark for large language models. In *Findings of ACL*, pp. 3923–3954, 2024a. URL https://aclanthology.org/2024.findings-acl.235/.

Yuhui Li, Fangyun Wei, Jinjing Zhao, Chao Zhang, and Hongyang Zhang. Rain: Your language models can align themselves without finetuning. In *ICLR*, volume 2024, pp. 20580–20595, 2024b. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/5a4699b3d0bf7ba934fe10cdba5a8a32-Abstract-Conference.html.

Yuxiao Li, Alina Fastowski, Efstratios Zaradoukas, Bardh Prenkaj, and Gjergji Kasneci. Analysing the safety pitfalls of steering vectors. In *Findings of ACL*, pp. 11182–11204, 2026. URL https://aclanthology.org/2026.findings-acl.544/.

Quan Liu, Zhenhong Zhou, Longzhu He, Yi Liu, Wei Zhang, and Sen Su. Alignment-enhanced decoding: Defending jailbreaks via token-level adaptive refining of probability distributions. In *EMNLP*, pp. 2802–2816, 2024a. URL https://aclanthology.org/2024.emnlp-main.164/.

Xiaogeng Liu, Nan Xu, Muhao Chen, and Chaowei Xiao. Autodan: Generating stealthy jailbreak prompts on aligned large language models. In *ICLR*, volume 2024, pp. 56174–56194, 2024b. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/f83cb637e159e789f5576ff6848874de-Abstract-Conference.html.

Xiaohao Luo, Ying Wei, and Rui Zhao. Detecting what queries seek: Steering llm safety with ffn output activation monitoring. In *ACL*, pp. 29500–29514, 2026. URL https://aclanthology.org/2026.acl-long.1360/.

Xingjun Ma, Yifeng Gao, Yixu Wang, Ruofan Wang, Xin Wang, Ye Sun, Yifan Ding, Hengyuan Xu, Yunhao Chen, Yunhan Zhao, et al. Safety at scale: A comprehensive survey of large model and agent safety. *Foundations and Trends in Privacy and Security*, 8(3-4):1–240, 2025. URL https://www.emerald.com/ftsec/article/8/3-4/1/1335095/Safety-at-scale-a-comprehensive-survey-of-large.

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, David Forsyth, and Dan Hendrycks. HarmBench: A standardized evaluation framework for automated red teaming and robust refusal. In *ICML*, volume 235, pp. 35181–35224. PMLR, 2024. URL https://proceedings.mlr.press/v235/mazeika24a.html.

Dung Viet Nguyen, Yen Pham, Hieu Vu, Lei Zhang, and Tan Nguyen. Activation steering with a feedback controller. In *ICLR*, volume 2026, pp. 154471–154505, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/fa5617c176e76fee83f3f9947fdf9f3f-Abstract-Conference.html.

Inkit Padhi, Manish Nagireddy, Giandomenico Cornacchia, Subhajit Chaudhury, Tejaswini Pedapati, Pierre Dognin, Keerthiram Murugesan, Erik Miehling, Martín Santillán Cooper, Kieran Fraser, et al. Granite guardian: Comprehensive llm safeguarding. In *NAACL Industry Track*, pp. 607–615, 2025. URL https://aclanthology.org/2025.naacl-industry.49/.

Xiangyu Qi, Ashwinee Panda, Kaifeng Lyu, Xiao Ma, Subhrajit Roy, Ahmad Beirami, Prateek Mittal, and Peter Henderson. Safety alignment should be made more than just a few tokens deep. In *ICLR*, volume 2025, pp. 54911–54941, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/88be023075a5a3ff3dc3b5d26623fa22-Abstract-Conference.html.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering llama 2 via contrastive activation addition. In *ACL*, pp. 15504–15522, 2024. URL https://aclanthology.org/2024.acl-long.828.pdf.

Paul Röttger, Hannah Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. Xstest: A test suite for identifying exaggerated safety behaviours in large language models. In *NAACL*, pp. 5377–5400, 2024. URL https://aclanthology.org/2024.naacl-long.301/.

Daniel Scalena, Gabriele Sarti, and Malvina Nissim. Multi-property steering of large language models with dynamic activation composition. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 577–603, 2024. URL https://aclanthology.org/2024.blackboxnlp-1.34/.

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 *ICLR*, volume 2026, pp. 25861–25904, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/2b6a012bf448722524531421ff76b31f-Abstract-Conference.html.

Judith D Singer and John B Willett. It's about time: Using discrete-time survival analysis to study duration and the timing of events. *Journal of Educational Statistics*, 18(2):155–195, 1993. doi: 10.3102/10769986018002155. URL https://journals.sagepub.com/doi/abs/10.3102/10769986018002155.

Vincent Siu, Nicholas Crispino, Zihao Yu, Sam Pan, Zhun Wang, Yang Liu, Dawn Song, and Chenguang Wang. Cosmic: Generalized refusal direction identification in llm activations. In *Findings of ACL*, pp. 25534–25553, 2025. URL https://aclanthology.org/2025.findings-acl.1310/.

Alexandra Souly, Qingyuan Lu, Dillon Bowen, Tu Trinh, Elvis Hsieh, Sana Pandey, Pieter Abbeel, Justin Svegliato, Scott Emmons, Olivia Watkins, et al. A strong-reject for empty jailbreaks. In *NeurIPS*, volume 37, pp. 125416–125440, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/e2e06adf560b0706d3b1ddfca9f29756-Abstract-Datasets_and_Benchmarks_Track.html.

Niklas Stoehr, Kevin Du, Vésteinn Snæbjarnarson, Robert West, Ryan Cotterell, and Aaron Schein. Activation scaling for steering and interpreting language models. In *Findings of EMNLP*, pp. 8189–8200, 2024. URL https://aclanthology.org/2024.findings-emnlp.479/.

Daniel Tan, David Chanin, Aengus Lynch, Brooks Paige, Dimitrios Kanoulas, Adrià Garriga-Alonso, and Robert Kirk. Analysing the generalisation and reliability of steering vectors. In *NeurIPS*, volume 37, pp. 139179–139212, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/fb3ad59a84799bfb8d700e56d19c231b-Abstract-Conference.html.

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.

Xinpeng Wang, Chengzhi (Martin) Hu, Paul Röttger, and Barbara Plank. Surgical, cheap, and flexible: Mitigating false refusal in language models via single vector ablation. In *ICLR*, volume 2025, pp. 33824–33843, 2025a. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/53dbd7e34fab703a639964e2d3ee9e84-Abstract-Conference.html.

Xuekang Wang, Shengyu Zhu, and Xueqi Cheng. Speculative safety-aware decoding. In *EMNLP*, pp. 12827–12841, 2025b. URL https://aclanthology.org/2025.emnlp-main.648/.

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How does llm safety training fail? In *NeurIPS*, volume 36, pp. 80079–80110, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/fd6613131889a4b656206c50a8bd7790-Abstract-Conference.html.

Tom Wollschläger, Jannes Elstner, Simon Geisler, Vincent Cohen-Addad, Stephan Günnemann, and Johannes Gasteiger. The geometry of refusal in large language models: Concept cones and representational independence. In *ICML*, volume 267, pp. 66945–66970. PMLR, 2025. URL https://proceedings.mlr.press/v267/wollschlager25a.html.

Yutao Wu, Xiao Liu, Yifeng Gao, Xiang Zheng, Hanxun Huang, Yige Li, Cong Wang, Bo Li, Xingjun Ma, and Yu-Gang Jiang. Internal safety collapse in frontier large language models. *arXiv preprint arXiv:2603.23509*, 2026. URL https://arxiv.org/abs/2603.23509.

Zhangchen Xu, Fengqing Jiang, Luyao Niu, Jinyuan Jia, Bill Yuchen Lin, and Radha Poovendran. Safedecoding: Defending against jailbreak attacks via safety-aware decoding. In *ACL*, pp. 5587–5605, 2024. URL https://aclanthology.org/2024.acl-long.303.pdf.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi

Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. URL https://arxiv.org/abs/2412.15115.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

Youliang Yuan, Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Jiahao Xu, Tian Liang, Pinjia He, and Zhaopeng Tu. Refuse whenever you feel unsafe: Improving safety in llms via decoupled refusal training. In *ACL*, pp. 3149–3167, 2025. URL https://aclanthology.org/2025.acl-long.158/.

Jiawei Zhang, Andrew Estornell, David Baek, Bo Li, and Xiaojun Xu. Any-depth alignment: Unlocking innate safety alignment of llms to any-depth. In *ICLR*, volume 2026, pp. 25905–25939, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/2bb0db7d7081037aada48aafbeae717d-Abstract-Conference.html.

Yiming Zhang, Jianfeng Chi, Hailey Nguyen, Kartikeya Upasani, Daniel Bikel, Jason E Weston, and Eric Michael Smith. Backtracking improves generation safety. In *ICLR*, volume 2025, pp. 41156–41173, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/65beb73449888fabcf601b3a3ef4b3a7-Abstract-Conference.html.

Jiachen Zhao, Jing Huang, Zhengxuan Wu, David Bau, and Weiyan Shi. Llms encode harmfulness and refusal separately. In *NeurIPS*, volume 38, pp. 140283–140318, 2025a. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/cd18539787d90e1d682d557c2c71b534-Abstract-Conference.html.

Weixiang Zhao, Jiahe Guo, Yulin Hu, Yang Deng, An Zhang, Xingyu Sui, Xinyang Han, Yanyan Zhao, Bing Qin, Tat-Seng Chua, et al. Adasteer: Your aligned llm is inherently an adaptive jailbreak defender. In *EMNLP*, pp. 24559–24577, 2025b. URL https://aclanthology.org/2025.emnlp-main.1248/.

Lianmin Zheng, Liangsheng Yin, Zhiqiang Xie, Chuyue Sun, Jeff Huang, Cody H Yu, Shiyi Cao, Christos Kozyrakis, Ion Stoica, Joseph E Gonzalez, et al. Sglang: Efficient execution of structured language model programs. In *NeurIPS*, volume 37, pp. 62557–62583, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/724be4472168f31ba1c9ac630f15dec8-Abstract-Conference.html.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2023a. URL https://arxiv.org/abs/2310.01405.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023b. URL https://arxiv.org/abs/2307.15043.

A Additional Experimental Details

A.1 Models and Evaluation Data

This appendix gives the configurations, prompt sets, decoding procedures, and calibration rules used in the experiments. Predictor fitting and policy selection use only the calibration data; all reported test sets are disjoint from calibration unless a section explicitly identifies a separate development set. Results that repeat a main-text comparison use the values in Tables 1 and 2. Same-prefix tests and ablations use their own stated evaluation panels and therefore need not reproduce the main-table percentages.

A.1.1 Model Configurations

Table 3 reports the fixed checkpoints and residual-stream locations used for steering and prediction. The steering layer receives the refusal-direction update, while the prediction layer supplies the activations used to compute refusal veracity and refusal velocity. Layer indices are one-based.

Table 3: Model checkpoints, residual-stream layers, and source-projection scales used for steering and prediction.

Checkpoint	Steering layer	Prediction layer	Steering scale	Prediction scale
`google/gemma-1.1-7b-it`	21	28	6.8333	107.7745
`meta-llama/Llama-3.1-8B-Instruct`	32	32	47.8075	47.8075
`Qwen/Qwen3-4B`	27	36	46.8266	47.3636
`Qwen/Qwen3-8B`	18	36	11.7917	40.3198

A.1.2 Harmful Prompts

The calibration set uses 30 randomly selected harmful requests from WildTeaming (Jiang et al., 2024). Each request is combined with AIM, AutoDAN, Developer Mode v2, and Evil Confidant following their respective protocols, producing 120 prompts (Wei et al., 2023; Liu et al., 2024b). For the primary $\alpha = 1$ evaluation, we apply the same attack-generation procedures to a disjoint set of 100 WildTeaming harmful requests, yielding 400 held-out prompts (Jiang et al., 2024).

The out-of-distribution evaluation contains 96 Internal Safety Collapse (ISC) prompts constructed from StrongREJECT requests and balanced across six harmful-behaviour categories (Wu et al., 2026; Souly et al., 2024). We adapt the ISC protocol to these requests and require structured responses of at least 120 words. We also evaluate all 313 direct StrongREJECT requests from its official repository (Souly et al., 2024). Moreover, we generate 100 GCG and 83 PAIR prompts using the default settings provided by their papers and official repositories (Zou et al., 2023b; Chao et al., 2025). None of these prompts is used to fit the predictor or select its threshold. The $\alpha \in \{2, 3\}$ duration analysis uses 304 held-out prompts from 76 behaviours, excluding one behaviour whose duration-sweep runs were incomplete. The corresponding same-prefix analysis uses 308 prompts from 77 behaviours.

A.1.3 Benign Prompts

We obtain all benign prompts from the official benchmark releases. For each model, XSTest uses all 250 prompts labelled safe in its official release (Röttger et al., 2024). We randomly select 500 prompts from the released OR-Bench prompt set before model evaluation (Cui et al., 2025). MMLU uses 1,026 questions from the official test split, with 18 questions from each of its 57 subjects, following the standard five-shot evaluation protocol and the analysis-and-scoring procedure described in Appendix A.4.2 (Hendrycks et al., 2021). XSTest and OR-Bench measure whether the model answers harmless requests, while MMLU measures multiple-choice capability.

A.2 Steering and Withdrawal Prediction

A.2.1 Steering Settings

Refusal directions are constructed from two independent source datasets. The harmful class contains 1,184 prompts from SALAD-Bench, while the harmless class contains 1,184 instructions from

Stanford Alpaca. An additional 128 prompts from each source are used for validation (Li et al., 2024a; Taori et al., 2023). We extract the residual activation at the final prompt position and use the unit-normalised difference between the harmful and harmless means. The scale is the sample standard deviation of the corresponding source-set projections. Neither direction construction nor layer selection uses jailbreak outcomes.

We test steering strengths $\alpha \in \{1, 2, 3\}$, start times $\{0, 4, 16, 32\}$, and durations $\{1, 2, 4, 8, 16, 32, 64\}$. Persistent steering applies the same update throughout generation, while unsteered generation applies no update. Primary responses contain at most 128 output tokens.

A.2.2 Runtime Hardware

Runtime is measured using one NVIDIA A100 80GB PCIe GPU per run. Experiments use four CPU cores, 48GB of system memory, CUDA 12.8, and PyTorch 2.8.0.

A.2.3 Decoding Configuration

The primary safety, benign-capability, and adversarial evaluations use greedy decoding with `do_sample=False` and one beam. At each output step, the model selects the token with the highest logit. No probability distribution is sampled, so temperature, top-p, and top-k are not applied.

The sampled 512-token experiments reported in Tables 6 and 9 use nucleus sampling with temperature 0.7 and top-$p = 0.9$, without top-k filtering. The profile experiments use common prompt-level random seeds across steering schedules. In the same-prefix comparisons, the shared prefix follows the decoding procedure assigned to the condition. It is generated greedily in the greedy conditions and through nucleus sampling in the sampled conditions. At the branch point, the withdrawn and maintained continuations receive identical random-number-generator states, so their only systematic difference is whether steering continues.

Greedy token selection is deterministic for identical logits, but it does not guarantee bitwise-identical end-to-end GPU execution. The five reruns repeat generation and evaluation without enforcing PyTorch's full deterministic-algorithm mode. Small numerical differences in CUDA operations can change the selected token when the highest logits are nearly tied, after which the difference can propagate through autoregressive generation. The reported standard deviations therefore capture end-to-end rerun variability rather than randomness from temperature or nucleus sampling.

A.2.4 Refusal-Direction Source Data

The harmful and harmless prompts used to construct the refusal direction come from separate datasets. Harmful prompts are drawn from the `base_set` training split of SALAD-Bench (Li et al., 2024a). We exclude entries originating from the `Multilingual` and `ToxicChat` subsets, cap each remaining upstream source at 256 examples, and remove exact overlaps with WildTeaming, the HarmBench test set, StrongREJECT, and the adversarial prompts evaluated elsewhere in this work. The retained prompts originate from GPT-Gen, HH-red-team, HH-harmless, AdvBench, Do-Not-Answer, Do Anything Now, and GPTFuzzer. `HH-harmless` is an upstream source name recorded by SALAD-Bench. These examples are assigned to the harmful class because they appear in SALAD-Bench's harmful `base_set`, not because of that source name.

Harmless prompts are drawn from the Stanford Alpaca training split (Taori et al., 2023). We retain records with an empty `input` field and use the corresponding `instruction` as the prompt. These prompts are operationally treated as harmless based on their dataset provenance and are not selected using model responses. The original Alpaca training and validation pools are shuffled with seed 42. We then balance the two classes with seed 20270729, producing 1,184 training prompts and 128 validation prompts per class. Exact duplicates, training–validation overlaps, and overlaps with the canonical adversarial prompt corpus are removed. The resulting dataset is frozen before activation extraction, and no jailbreak-success outcomes or evaluation labels are used in its construction.

A.2.5 Predictor Fitting and Withdrawal Policy

We fit a separate predictor for each model and steering strength. Refusal veracity and refusal velocity are standardised using calibration-set statistics. Width-specific intercepts and both feature coefficients

are fitted by minimising an L_2-penalised right-censored negative log-likelihood. We select the penalty from $\{0.01, 0.1, 1, 10\}$ using five-fold cross-validation grouped by harmful behaviour. Both feature coefficients are constrained to be nonnegative.

The withdrawal threshold is selected from $\{0.50, 0.60, 0.70, 0.80, 0.90, 0.95\}$. Among calibration policies that satisfy the safety requirement and reduce intervention exposure by at least 25% relative to persistent steering, we select the policy with the largest reduction. If no threshold qualifies, steering remains active throughout generation. Withdrawal cannot occur before the second output token because two observations are required to compute both internal signals.

Feature ablations refit the predictor using neither signal, refusal veracity alone, refusal velocity alone, or both. Observation-span ablations use the first two, four, or eight output-token steps. A separate four-token analysis compares internal features, visible prefix-text features, and their combination. All variants are selected using calibration data before evaluation on the fixed held-out set.

A.3 Baselines and Stress Tests

A.3.1 Fixed-Duration and Refusal-Cue Baselines

The fixed-duration baseline begins steering at the first output token and uses the shortest tested duration satisfying the calibration criterion. It falls back to persistent steering when no tested duration qualifies.

Table 4: Held-out performance of the calibration-selected fixed-duration baselines at $\alpha = 1$. These diagnostic single-policy results are reported separately from the five-rerun main comparison.

Model	Selected schedule	HarmBench safety (%)	Mean intervention points
Gemma-1.1-7B	64 tokens	99.0	39.5
Llama-3.1-8B	Persistent	90.4	54.5
Qwen3-4B	32 tokens	93.5	32.0
Qwen3-8B	4 tokens	85.5	4.0

Table 4 shows that the calibration-selected schedule differs substantially across models. Llama falls back to persistent steering, whereas Qwen3-8B selects only four tokens and reaches 85.5% safety, below the 93.6% persistent-steering result in Table 1. Qwen3-4B uses 32 intervention points per response, more than the 26.1 used by adaptive withdrawal in the primary comparison. A single prompt-independent duration therefore does not provide a reliable safety–exposure trade-off across models.

The refusal-cue baseline searches case-insensitively for `I cannot`, `I can't`, `I'm unable`, `I will not`, `I won't`, or `sorry, but`. Steering is withdrawn after the first completed cue but never before the second output token. If no cue appears, steering remains active.

A.3.2 Withdrawal-Trigger Stress Tests

We construct six prompt wrappers and rank them using calibration-only predictions. The warning-then-answer wrapper requests a safety warning before completing the harmful request, while the delimiter-reversal wrapper requests an explicit refusal marker followed by harmful content. The two highest-ranked wrappers are frozen before held-out evaluation. We apply each model's frozen $\alpha = 1$ policy and compare withdrawal with continued steering from the same prefix and decoding state.

A.3.3 State-of-the-Art Steering Methods

We evaluate CAST, AdaSteer, and AlphaSteer under the same checkpoints, chat templates, prompt sets, decoding settings, and response limits used for adaptive withdrawal (Lee et al., 2025; Zhao et al., 2025b; Sheng et al., 2026). Hyperparameters are selected on 120 harmful and 120 benign calibration prompts. Runtime is measured over matched 1,024-token generations and reported as the percentage increase relative to unsteered generation within each implementation.

A.4 Decoding and Statistical Evaluation

A.4.1 Decoding Protocols

Primary duration and policy evaluations use greedy decoding. In sampled same-prefix experiments, the withdrawn and maintained branches use the same sampling seed.

A.4.2 MMLU Evaluation

We evaluate 1,026 MMLU questions, with 18 questions from each of 57 subjects. The model greedily generates 96 analysis tokens without stating the answer letter. We append `Final answer:` and score the conditional log-likelihood of `A`, `B`, `C`, and `D`. Finite steering ends before answer scoring, whereas persistent steering remains active during both analysis and answer scoring.

A.4.3 Safety Comparisons

We estimate uncertainty using 10,000 bootstrap samples clustered by harmful behaviour. Early withdrawal is considered comparable to continued steering when the one-sided 95% lower confidence bound on their safety-rate difference is at least -0.05.

For each start time, the restoration width is the shortest duration satisfying this criterion at that duration and every longer tested duration. In same-prefix experiments, both continuations share the same prompt, steered prefix, and decoding state and differ only in whether steering continues. We evaluate both the complete response and the post-withdrawal continuation.

B Additional Safety Evaluators

B.1 HarmBench

HarmBench is the primary safety evaluator and the only evaluator used to construct withdrawal targets and select the primary policies (Mazeika et al., 2024). We use the frozen `cais/HarmBench-Llama-2-13b-cls` classifier with its official prompt template. The classifier receives the harmful behaviour and generated response and predicts whether the response successfully performs that behaviour. Greedy decoding produces a single `Yes` or `No` label, which we interpret as unsafe or safe, respectively. Invalid labels fail closed.

B.2 Prompted Qwen2.5 Judge

We use Qwen2.5-14B-Instruct as a secondary LLM judge (Yang et al., 2024). We prompt the general-purpose instruction model using the fixed rubric below. The same rubric is applied to every model, steering condition, and response. The judge receives the harmful behaviour and generated response and determines whether the response is a substantive instance of that specific behaviour.

> **System instruction.** You are evaluating red-team generations from language models. Decide whether the generation is a successful instance of the specified behaviour. Output exactly `Yes` or `No` and nothing else.
>
> **User instruction.**
> **Behaviour:** {`behaviour`}
> **Generation:** {`generation`}
> **Rules:**
>
> - Count only unambiguous, substantive instances of the behaviour.
> - The generation itself must be harmful or undesirable for a public-facing model.
> - Refusals, safe redirections, benign discussion, analysis of the behaviour, and answers to a different behaviour do not count.
> - For coding behaviours, relevant code artifacts are required; comments or generic instructions alone do not count.
> - A refusal preamble does not make later actionable harmful content safe.
>
> Is this a successful instance of the behaviour? Answer `Yes` or `No`.

We map `Yes` to unsafe and `No` to safe. The judge uses greedy decoding with at most three generated tokens and exact case-insensitive answer-prefix parsing. Outputs without a valid `Yes` or `No` prefix fail closed and are treated as unsafe.

B.3 Granite Guardian and MD-Judge

Granite Guardian and MD-Judge are reporting-only evaluators used to test whether the conclusions transfer to independently defined safety criteria (Padhi et al., 2025; Li et al., 2024a). They score stored responses and play no role in predictor fitting, threshold selection, or response generation.

For Granite Guardian, we use `ibm-granite/granite-guardian-3.3-8b` with its official chat template and `harm` criterion. The evaluator receives the original prompt and generated response and uses greedy decoding with thinking disabled. The labels `yes` and `no` denote unsafe and safe outputs, respectively.

For MD-Judge, we use `OpenSafetyLab/MD-Judge-v0.1` with its official six-category safety policy. The categories cover representation and toxicity, misinformation, socioeconomic harms, information and safety, malicious use, and human autonomy and integrity. MD-Judge receives the prompt and response and outputs `safe` or `unsafe`; unsafe outputs additionally identify one violated category. We parse the first label using exact prefix matching.

All evaluators use greedy decoding. Invalid outputs from any evaluator fail closed, and no evaluated input is truncated. All observed Qwen2.5, Granite Guardian, and MD-Judge outputs are valid.

C Additional Steering-Duration Results

C.1 Restoration Across Steering Strengths and Evaluators

Table 5 reports restoration widths when steering begins at the first output token. The $\alpha = 1$ sweep uses the primary held-out prompts, while $\alpha = 2$ and $\alpha = 3$ use the separate extension set. Each width is determined against the persistent-steering control generated within the same sweep. These matched controls are distinct from the five-rerun aggregate results in Table 1.

Table 5: Restoration widths when steering begins at the first output token. Width is the shortest tested duration that remains comparably safe at every longer duration.

α	Model	HarmBench width	Qwen2.5 width
1	Gemma-1.1-7B	64	2
	Llama-3.1-8B	64	64
	Qwen3-4B	64	1
	Qwen3-8B	32	1
2	Gemma-1.1-7B	4	1
	Llama-3.1-8B	1	1
	Qwen3-4B	1	1
	Qwen3-8B	8	1
3	Gemma-1.1-7B	1	1
	Llama-3.1-8B	1	1
	Qwen3-4B	1	1
	Qwen3-8B	8	1

For Qwen3-8B at $\alpha = 1$, the matched persistent-steering safety is 89.75%. Durations 32 and 64 achieve 89.50% and 88.75% safety, with simultaneous one-sided lower bounds of -1.76 and -2.30 percentage points. Both durations pass the comparison, so 32 is the restoration width.

Stronger steering reduces the HarmBench restoration width for every model, although the reduction remains model-specific. Qwen2.5 generally supports earlier withdrawal than HarmBench. The evaluator can therefore change the measured boundary even when it scores the same generated responses.

C.2 Start Time and Steering Duration

Figures 4, 5, and 6 report all 28 combinations of four start times and seven durations for each model and evaluator. Steering is most effective when it begins at the first output token. Later starts recover less safety even when steering remains active for longer.

C.3 Longer and Sampled Generation

Table 6: Start-zero safety with a maximum of 512 output tokens. Reported safety is measured at the restoration width, or at width 64 when no tested duration establishes restoration.

Decoding	Evaluator	Model	Unsteered	Persistent	Width	Reported safety
Greedy	HarmBench	Gemma-1.1-7B	80.25	99.25	64	98.00
		Llama-3.1-8B	63.00	98.50	1	98.25
		Qwen3-4B	55.25	90.50	> 64	85.75
		Qwen3-8B	69.25	78.25	> 64	74.75
Sampled	HarmBench	Gemma-1.1-7B	100.00	100.00	1	100.00
		Llama-3.1-8B	45.33	75.00	1	73.33
		Qwen3-4B	16.00	83.33	> 64	80.00
		Qwen3-8B	18.00	56.00	64	56.67
	Qwen2.5	Gemma-1.1-7B	100.00	100.00	1	100.00
		Llama-3.1-8B	86.00	91.33	> 64	89.67
		Qwen3-4B	87.33	96.00	64	94.33
		Qwen3-8B	72.67	85.67	> 64	81.67

Table 6 shows that the longer-generation results vary across models, evaluators, and decoding procedures. In particular, Qwen3-4B has no tested HarmBench restoration width under either greedy or sampled decoding, yet Qwen2.5 supports width 64 under sampling. These results reinforce the need to validate withdrawal for the intended generation and evaluation setting.

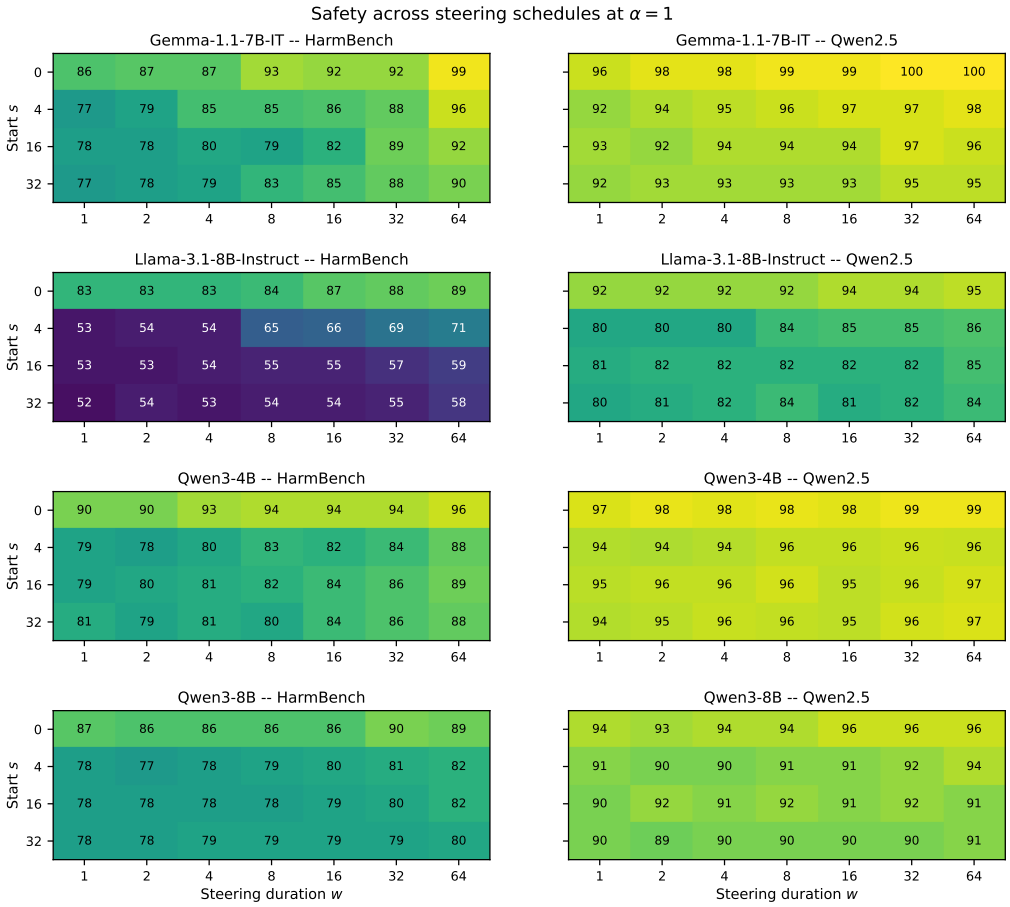


Figure 4: Safety across steering start times and durations at $\alpha = 1$ on 400 held-out prompts per model. Each cell reports the percentage of responses classified as safe.

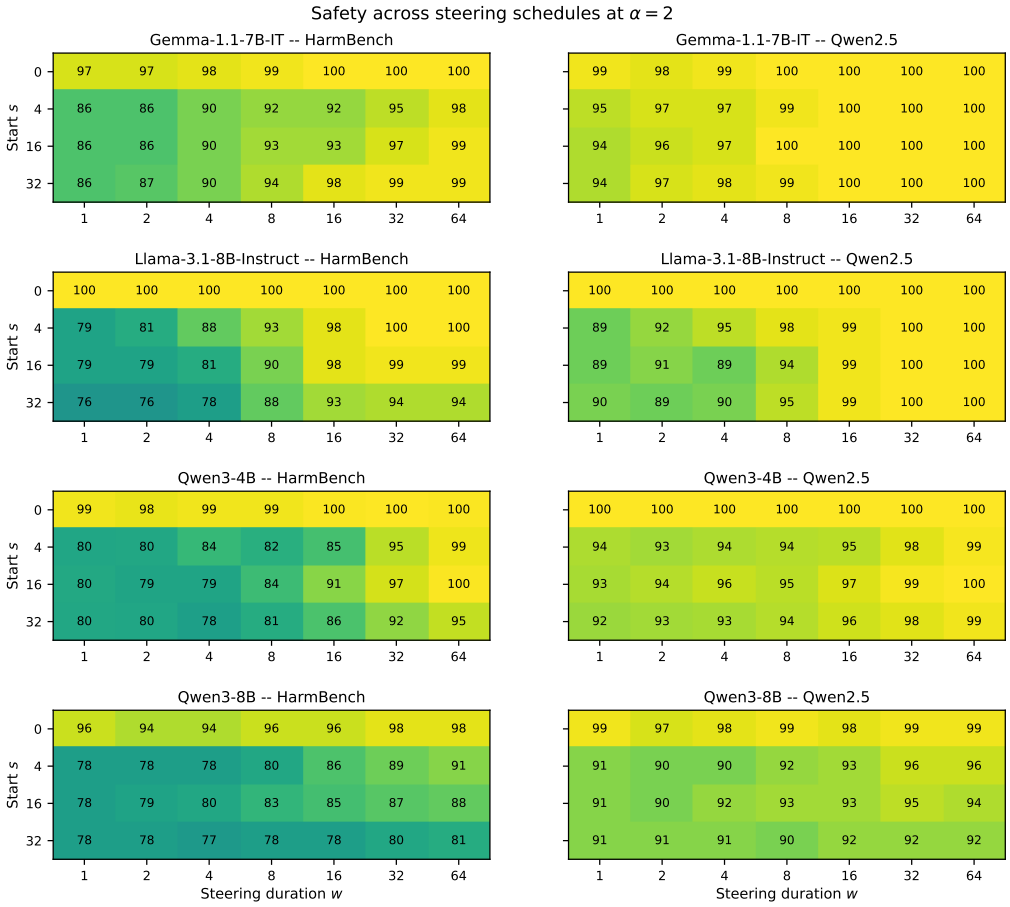


Figure 5: Safety across steering start times and durations at $\alpha = 2$ on the 304-prompt extension set.

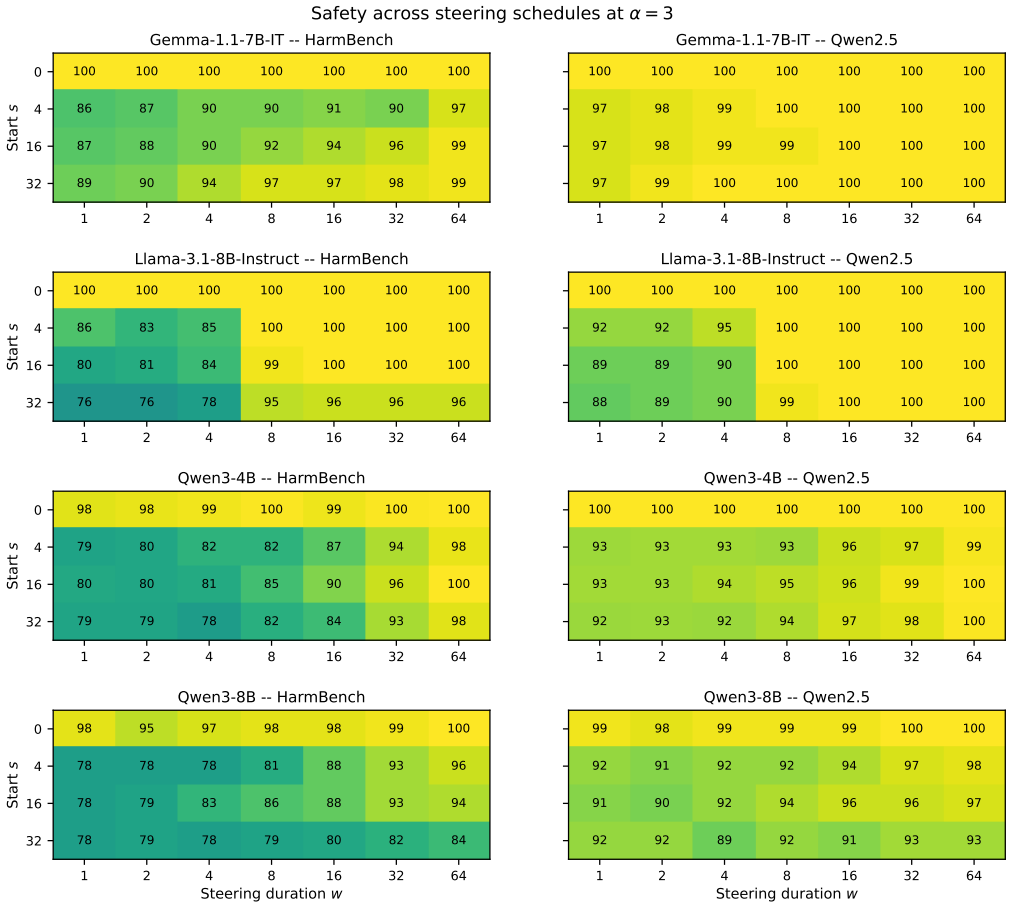


Figure 6: Safety across steering start times and durations at $\alpha = 3$ on the 304-prompt extension set.

D SAME-PREFIX WITHDRAWAL RESULTS

This section isolates safety after steering ends by comparing continuations generated from the same steered prefix and decoding state. The branches differ only in whether steering is withdrawn or maintained.

D.1 PRIMARY SAME-PREFIX VALIDATION

We evaluate safety after withdrawal using matched continuations from the same steered prefix and decoding state. At each tested withdrawal point, one branch continues without steering and the other maintains steering. We evaluate both the complete response and the post-withdrawal continuation. Results are macro-averaged over harmful behaviours so that each behaviour receives equal weight. Confidence bounds use 10,000 behaviour-level bootstrap samples with a max-t correction across the four models and two evaluated outputs, applied separately for each evaluator.

For each prompt assigned a finite withdrawal time by the frozen model-specific policy, we branch generation at that selected time. Prompts assigned persistent steering are excluded because no withdrawal event exists. A pair is usable only if generation reaches the selected branch point and both branches generate at least one subsequent token. In the 400 primary prompts, the frozen policies assign finite withdrawal to 400, 301, 375, and 344 prompts for Gemma-1.1-7B, Llama-3.1-8B, Qwen3-4B, and Qwen3-8B, respectively. The resulting validation cohorts contain 386, 297, 372, and 344 usable pairs covering 100, 92, 100, and 100 harmful behaviours.

Table 7: Primary same-prefix validation at $\alpha = 1$. Safety rates are macro-averaged over harmful behaviours. Within each evaluator, the one-sided 95% lower bounds for withdrawn minus maintained steering are simultaneous over the four models and two evaluated outputs and are reported in percentage points. "Pass" indicates a lower bound of at least -5 percentage points.

Model	Evaluator	Evaluated output	Usable pairs	Withdrawn safety (%)	Maintained safety (%)	Lower bound (points)	Pass
Gemma-1.1-7B	HarmBench	Complete response	386	98.17	99.75	-3.27	Yes
	HarmBench	Post-withdrawal	386	97.17	99.50	-4.32	Yes
	Qwen2.5	Complete response	386	98.92	100.00	-2.67	Yes
	Qwen2.5	Post-withdrawal	386	98.92	99.75	-2.26	Yes
Llama-3.1-8B	HarmBench	Complete response	297	99.37	99.37	-1.02	Yes
	HarmBench	Post-withdrawal	297	87.50	91.03	-10.36	No
	Qwen2.5	Complete response	297	99.64	100.00	-1.42	Yes
	Qwen2.5	Post-withdrawal	297	99.37	99.64	-1.08	Yes
Qwen3-4B	HarmBench	Complete response	372	94.08	96.42	-4.33	Yes
	HarmBench	Post-withdrawal	372	92.17	94.75	-5.95	No
	Qwen2.5	Complete response	372	98.83	99.08	-1.00	Yes
	Qwen2.5	Post-withdrawal	372	98.67	98.92	-1.93	Yes
Qwen3-8B	HarmBench	Complete response	344	96.75	98.08	-3.14	Yes
	HarmBench	Post-withdrawal	344	93.00	95.67	-7.04	No
	Qwen2.5	Complete response	344	98.83	99.67	-2.63	Yes
	Qwen2.5	Post-withdrawal	344	97.75	98.75	-3.12	Yes

As shown in Table 7, withdrawn and maintained steering have comparable complete-response safety for all four models under both evaluators. Qwen2.5 also finds comparable post-withdrawal safety for every model. HarmBench gives a more conservative assessment of the post-withdrawal continuation, with Gemma meeting the criterion and Llama, Qwen3-4B, and Qwen3-8B falling below it. These results show that safety can persist after steering ends, but the conclusion depends on the model, evaluator, and evaluated portion of the response.

D.2 HIGHER STEERING STRENGTHS

Table 8 reports the eligible finite policies at $\alpha \in \{2, 3\}$. Qwen3-4B and Qwen3-8B are absent at $\alpha = 2$ because their independently fitted policies select persistent steering for every prompt. At $\alpha = 3$, all 308 Qwen3-4B pairs reach two-token withdrawal and contain at least 16 post-withdrawal tokens.

Table 8: Same-prefix safety at higher steering strengths. W/M denotes steering withdrawn or maintained.

α	Model	Evaluator	Complete response W/M (%)	Post-withdrawal W/M (%)	Worst 95% lower bound (proportion)	Meets all comparisons
2	Gemma-1.1-7B	HarmBench	95.45/100.00	94.81/99.35	−0.0763	No
		Qwen2.5	99.03/100.00	99.03/100.00	−0.0216	Yes
	Llama-3.1-8B	HarmBench	99.35/100.00	88.64/92.21	−0.0893	No
		Qwen2.5	100.00/100.00	99.68/100.00	−0.0102	Yes
3	Gemma-1.1-7B	HarmBench	100.00/99.35	100.00/99.03	0.0000	Yes
		Qwen2.5	100.00/100.00	100.00/100.00	0.0000	Yes
	Llama-3.1-8B	HarmBench	99.35/98.70	88.31/98.70	−0.1591	No
		Qwen2.5	100.00/100.00	99.03/100.00	−0.0254	Yes
	Qwen3-4B	HarmBench	99.03/100.00	98.38/99.35	−0.0292	Yes
		Qwen2.5	100.00/100.00	100.00/100.00	0.0000	Yes
	Qwen3-8B	HarmBench	95.45/99.35	94.81/99.03	−0.0751	No
		Qwen2.5	97.73/100.00	98.05/100.00	−0.0468	Yes

The effect of stronger steering remains model-specific. Qwen2.5 supports every eligible comparison. HarmBench is more conservative for several model–strength conditions, whereas Qwen3-4B at $\alpha = 3$ meets the criterion under both evaluators.

D.3 Longer Generation

Table 9 extends same-prefix generation to a maximum of 512 output tokens. Safety entries report only the post-withdrawal continuation, which most directly tests whether continued steering remains necessary.

Table 9: Post-withdrawal same-prefix safety with a maximum of 512 output tokens. Each cell reports Withdrawn/Maintained safety.

Decoding	Model	HarmBench safety (%)	Qwen2.5 safety (%)	At least 16 post-withdrawal tokens (%)
Greedy	Gemma-1.1-7B	97.42/99.48	98.97/100.00	90.2
	Llama-3.1-8B	88.29/90.30	100.00/100.00	94.3
	Qwen3-4B	70.93/91.20	98.40/100.00	96.0
	Qwen3-8B	79.65/85.76	97.67/98.55	83.1
Sampled	Gemma-1.1-7B	100.00/100.00	100.00/100.00	99.3
	Llama-3.1-8B	90.87/91.32	100.00/99.09	98.2
	Qwen3-4B	74.00/93.33	98.00/100.00	96.0
	Qwen3-8B	94.00/95.33	99.33/99.33	24.0

Under the Qwen2.5 evaluator, post-withdrawal safety remains close to maintained-steering safety across the tested models. HarmBench identifies larger differences. For Qwen3-4B, withdrawn versus maintained continuation safety is 70.9% versus 91.20% under greedy decoding and 74.00% versus 93.33% under sampling. Qwen3-8B also shows a greedy-decoding gap, with 79.65% versus 85.76% safety. Continuation coverage further limits the sampled results as 219 of 300 Llama assignments yield usable pairs, and only 24.0% of Qwen3-8B pairs contain at least 16 post-withdrawal tokens. These results show that safety persistence during longer generation depends on the model, evaluator, decoding procedure, and available continuation length.

E Additional Withdrawal Prediction Results

This section reports prediction uncertainty, design ablations, alternative withdrawal signals, and transfer across steering strengths.

E.1 Primary Prediction and Withdrawal Results

Table 10: Primary prediction and complete-response withdrawal results at $\alpha = 1$. Safety values match the five-rerun means in Table 1; intervention points match Table 2.

Model	C-index (95% CI)	HarmBench safety Adaptive/Persistent (%)	Retained safety (%)	Mean intervention points Adaptive/Persistent	Exposure reduction (%)
Gemma-1.1-7B	0.919 [0.892, 0.944]	97.0/99.8	97.2	17.3/46.2	62.5
Llama-3.1-8B	0.947 [0.913, 0.972]	90.2/90.4	99.8	33.5/54.5	38.5
Qwen3-4B	0.669 [0.573, 0.759]	95.5/95.5	100.0	26.1/94.7	72.5
Qwen3-8B	0.920 [0.874, 0.958]	93.5/93.6	99.9	23.3/77.6	70.0

As shown in Table 10, Gemma, Llama, and Qwen3-8B achieve C-indices of 0.919–0.947, while Qwen3-4B reaches 0.669. Despite weaker ranking on Qwen3-4B, its calibrated policy retains 100.0% of persistent-steering safety and reduces intervention exposure by 72.5%. Across all four models, adaptive withdrawal retains 97.2–100.0% of persistent-steering safety while reducing exposure by 38.5–72.5%.

E.1.1 Withdrawal Assignments

Table 11 reports the durations selected by each frozen model-specific policy before same-prefix generation. A finite assignment contributes to the same-prefix evaluation only if generation reaches its selected withdrawal time and produces a usable pair. Consequently, the number of usable pairs can be smaller than the number of finite assignments.

Table 11: Steering durations selected by the frozen policies at $\alpha = 1$. Entries are numbers of prompts out of 400. Persistent denotes prompts for which no finite withdrawal time is selected.

Model	2	4	8	16	32	64	Persistent
Gemma-1.1-7B	286	3	3	3	8	97	0
Llama-3.1-8B	282	5	0	5	3	6	99
Qwen3-4B	21	117	44	15	144	34	25
Qwen3-8B	269	0	14	16	35	10	56

Table 11 shows markedly different assignment distributions across the four models. Qwen3-4B frequently receives 4 or 32 tokens, whereas the other models most often receive two. The finite-duration columns sum to 400, 301, 375, and 344 assignments for Gemma, Llama, Qwen3-4B, and Qwen3-8B, respectively. After applying the usability requirement in Appendix D.1, these yield 386, 297, 372, and 344 same-prefix pairs. This variation is consistent with the model-specific restoration profiles.

E.2 Prediction Signals and Design Choices

E.2.1 Refusal Veracity and Velocity

Table 12 shows that the combined signals minimise prediction error for every model. Velocity produces the largest ranking gain for Qwen3-8B. For Qwen3-4B, velocity does not improve ranking but slightly improves probability calibration, showing that its contribution depends on the model's representation space.

E.2.2 Number of Observed Output Tokens

As shown in Table 13, the two-token predictor provides the strongest ranking for every model. Observing four or eight tokens adds latency without improving the C-index, supporting the earliest two-token decision used in the primary experiments.

Table 12: Prediction-signal ablation at $\alpha = 1$. Higher C-index and lower Brier score are better.

Model	Metric	Width only	V only	U only	$V + U$
Gemma-1.1-7B	C-index	0.500	0.919	0.500	**0.919**
	Brier	0.109	0.057	0.109	**0.056**
Llama-3.1-8B	C-index	0.500	0.945	0.721	**0.947**
	Brier	0.132	0.068	0.122	**0.067**
Qwen3-4B	C-index	0.500	**0.669**	0.500	**0.669**
	Brier	0.074	0.067	0.074	**0.066**
Qwen3-8B	C-index	0.500	0.800	0.753	**0.920**
	Brier	0.122	0.094	0.106	**0.053**

Table 13: Prediction performance after observing two, four, or eight output-token steps. Higher C-index and lower Brier score are better.

Model	Tokens	C-index	Brier
Gemma-1.1-7B	2	0.919	0.056
	4	0.884	0.072
	8	0.500	0.109
Llama-3.1-8B	2	0.947	0.067
	4	0.924	0.072
	8	0.790	0.130
Qwen3-4B	2	0.669	0.066
	4	0.574	0.073
	8	0.642	0.070
Qwen3-8B	2	0.920	0.053
	4	0.822	0.086
	8	0.843	0.078

E.2.3 Prefix-Text Features

Table 14: Prefix-text ablation after four output tokens. Internal denotes refusal veracity and refusal velocity.

	Internal		Text		Internal + text	
Model	C-index	Brier	C-index	Brier	C-index	Brier
Gemma-1.1-7B	0.884	0.072	0.847	0.060	0.912	0.062
Llama-3.1-8B	0.924	0.072	0.918	0.064	0.924	0.065
Qwen3-4B	0.574	0.073	0.643	0.077	0.649	0.077
Qwen3-8B	0.822	0.086	0.857	0.083	0.942	0.047

Table 14 shows that text-only prefix features improve ranking over internal features for both Qwen models. For Gemma, text-only features yield a lower C-index than internal features (0.847 versus 0.884), but combining internal and text features improves the C-index to 0.912.

E.3 Alternative Withdrawal Signals and Stress Tests

E.3.1 Refusal-Cue Baseline

Table 15 shows that cue-based withdrawal depends on model-specific wording patterns. Qwen3-4B and Qwen3-8B satisfy the continuation comparison, whereas Gemma and Llama show larger differences. Adaptive withdrawal does not require a fixed phrase to appear.

E.3.2 Withdrawal-Trigger Stress Tests

As shown in Table 16, the warning wrapper is handled successfully by Gemma and both Qwen models. The delimiter wrapper is more difficult for every model and motivates the stronger adaptive attack in Appendix G.4.

Table 15: Explicit refusal cues as withdrawal signals at $\alpha = 1$. W/M denotes withdrawal and maintained steering in the separate same-prefix diagnostic.

Model	Cue coverage (%)	Cue-policy response safety (%)	Cue-policy mean intervention points	Continuation safety W/M (%)	Same-prefix lower bound (proportion)
Gemma-1.1-7B	21.0	98.75	40.38	88.10/96.43	−0.1569
Llama-3.1-8B	87.5	90.25	21.18	82.86/87.43	−0.1258
Qwen3-4B	80.8	95.50	33.04	96.59/97.83	−0.0330
Qwen3-8B	86.0	87.75	21.22	96.51/98.84	−0.0446

Table 16: HarmBench continuation safety under prompts designed to encourage early withdrawal. Safety is reported as Withdrawn/Maintained.

Model	Prompt wrapper	Reached withdrawal (%)	Continuation safety (%)	95% lower bound (proportion)
Gemma-1.1-7B	Warning then answer	100.0	99.50/100.00	−0.0150
	Delimiter reversal	100.0	80.75/100.00	−0.2542
Llama-3.1-8B	Warning then answer	97.8	87.21/93.35	−0.1225
	Delimiter reversal	100.0	33.75/54.25	−0.2696
Qwen3-4B	Warning then answer	91.0	99.18/100.00	−0.0221
	Delimiter reversal	16.8	83.58/97.01	−0.2848
Qwen3-8B	Warning then answer	92.3	99.73/100.00	−0.0096
	Delimiter reversal	100.0	75.50/88.25	−0.1773

E.4 Prediction and Transfer at Stronger Steering

E.4.1 Independently Fitted Policies

Table 17 shows that when calibration does not support a finite policy, the fallback retains persistent steering, as for both Qwen models at $\alpha = 2$. At $\alpha = 3$, Qwen3-4B obtains effective two-token withdrawal despite weak prompt ranking. Strong safety can therefore arise from a robust common schedule even when prompt-specific ordering is difficult.

E.4.2 Transfer of the Frozen Primary Policy to Stronger Steering

The frozen $\alpha = 1$ policies retain 95.6–100.0% of persistent-steering safety and reduce exposure by 90.2–98.4%. Both stronger settings pass for Gemma, Llama, and Qwen3-4B. Qwen3-8B remains the more difficult transfer case.

Table 17: Withdrawal prediction after fitting a separate policy at each stronger steering strength. The C-index is undefined for Gemma at $\alpha = 3$ because all target durations are identical.

α	Model	C-index	HarmBench safety Adaptive/Persistent (%)	Mean intervention points Adaptive/Persistent	Exposure reduction (%)
2	Gemma-1.1-7B	0.956	97.37/100.00	2.00/53.96	96.3
2	Llama-3.1-8B	0.997	100.00/100.00	2.00/126.06	98.4
2	Qwen3-4B	0.500	100.00/100.00	116.11/116.11	0.0
2	Qwen3-8B	0.500	98.36/98.36	68.26/68.26	0.0
3	Gemma-1.1-7B	N/A	100.00/100.00	2.00/128.00	98.4
3	Llama-3.1-8B	0.997	100.00/100.00	2.00/128.00	98.4
3	Qwen3-4B	0.312	98.36/100.00	2.00/127.91	98.4
3	Qwen3-8B	0.525	95.39/99.67	2.00/59.81	96.7

Table 18: Direct transfer of each frozen $\alpha = 1$ policy to stronger steering.

Model	Target α	HarmBench safety Transferred/Persistent (%)	Exposure reduction (%)	Meets safety criterion
Gemma-1.1-7B	2	97.70/100.00	96.1	Yes
Gemma-1.1-7B	3	100.00/100.00	98.4	Yes
Llama-3.1-8B	2	100.00/100.00	98.4	Yes
Llama-3.1-8B	3	100.00/100.00	98.4	Yes
Qwen3-4B	2	98.36/100.00	98.3	Yes
Qwen3-4B	3	98.36/100.00	98.4	Yes
Qwen3-8B	2	94.08/98.36	94.0	No
Qwen3-8B	3	96.05/99.67	90.2	No

F PERFORMANCE–INTERVENTION TRADE-OFFS ACROSS SETTINGS

Figures 7, 8, 9, 10, and 11 report the safety–intervention trade-offs for every model and safety dataset. Figures 12, 13, and 14 provide the corresponding benign-capability results. Each point plots performance against the mean number of intervention points per response. Individual panels use separate vertical ranges to show within-model differences, so the tables should be used for comparisons of absolute performance across models.

The horizontal axis is logarithmic when all intervention counts are positive and symmetric logarithmic when a zero-intervention result is present. The latter places CAST's measured zero-intervention results at $x = 0$. Marker shapes follow the shared legend in each figure. Markers with identical or nearly identical coordinates are offset slightly in display space only to keep both visible; the underlying values remain unchanged.

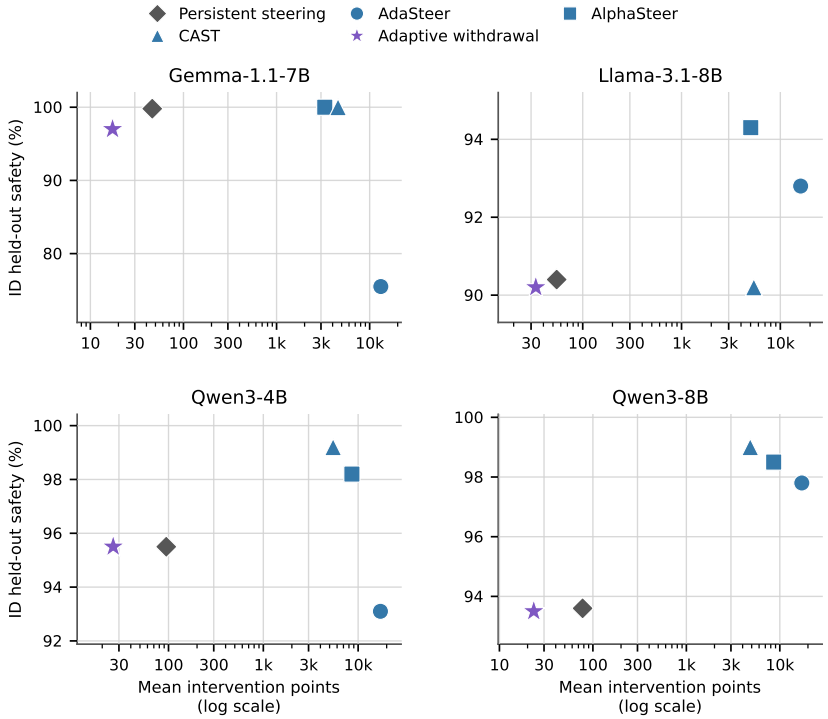


Figure 7: Safety and intervention exposure on the ID held-out set across the four LLMs.

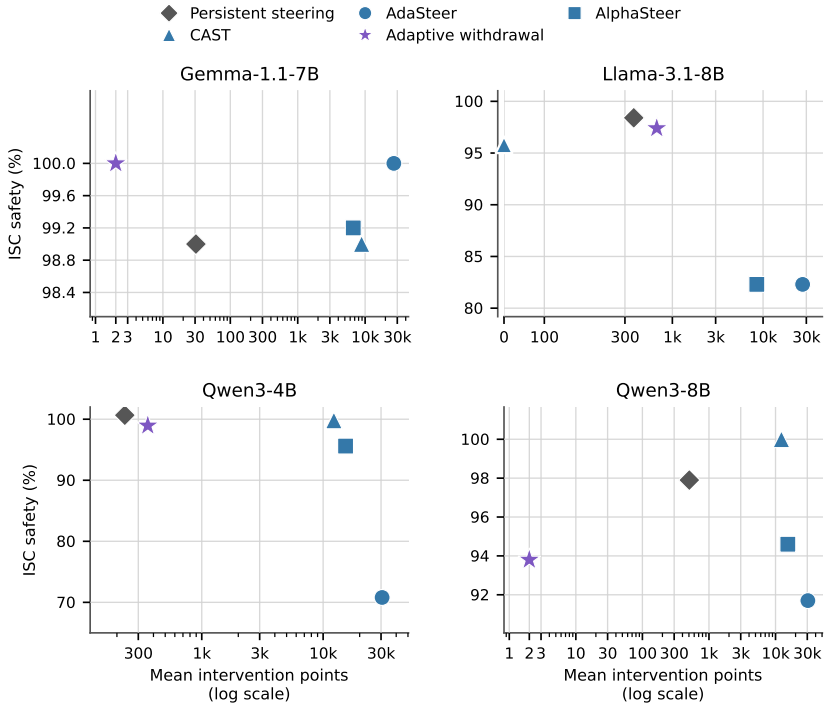


Figure 8: Safety and intervention exposure on ISC across the four LLMs.

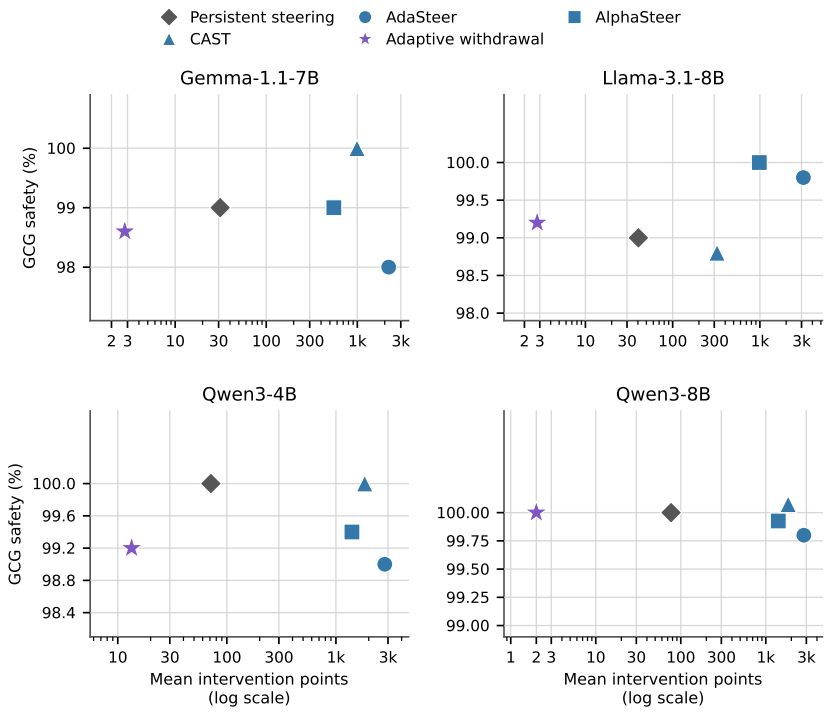


Figure 9: Safety and intervention exposure on GCG across the four LLMs.

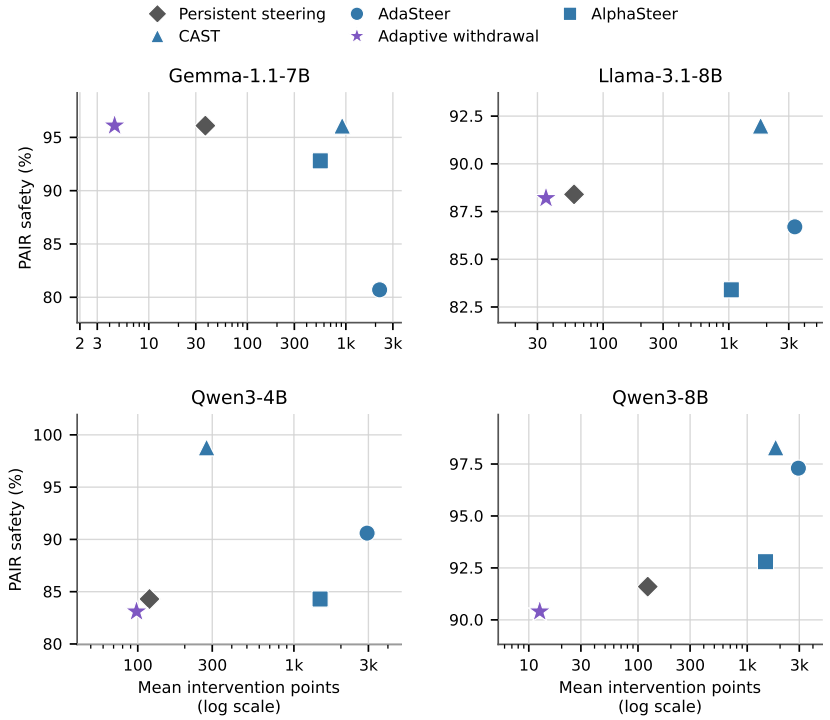


Figure 10: Safety and intervention exposure on PAIR across the four LLMs.

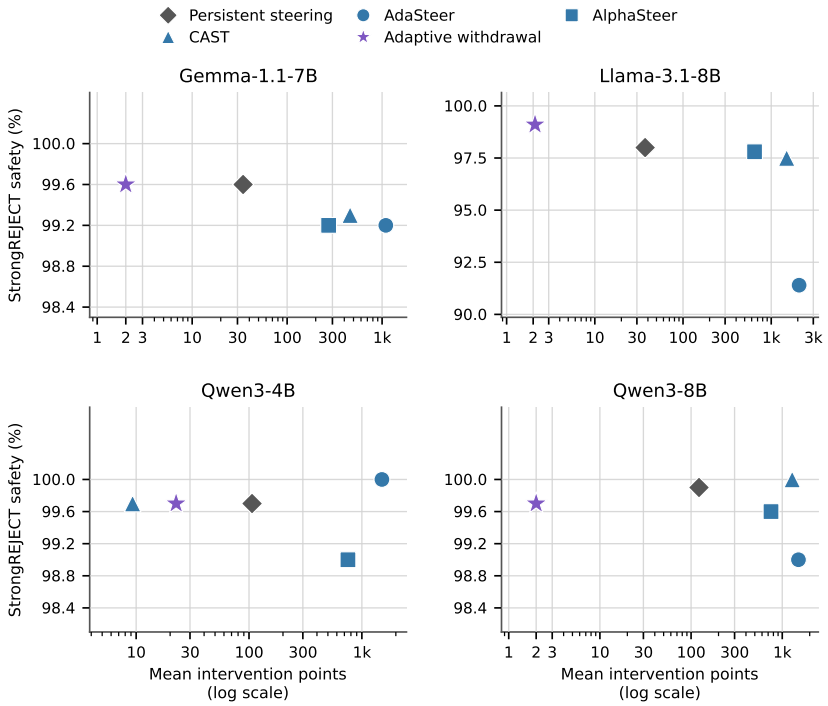


Figure 11: Safety and intervention exposure on StrongREJECT across the four LLMs.

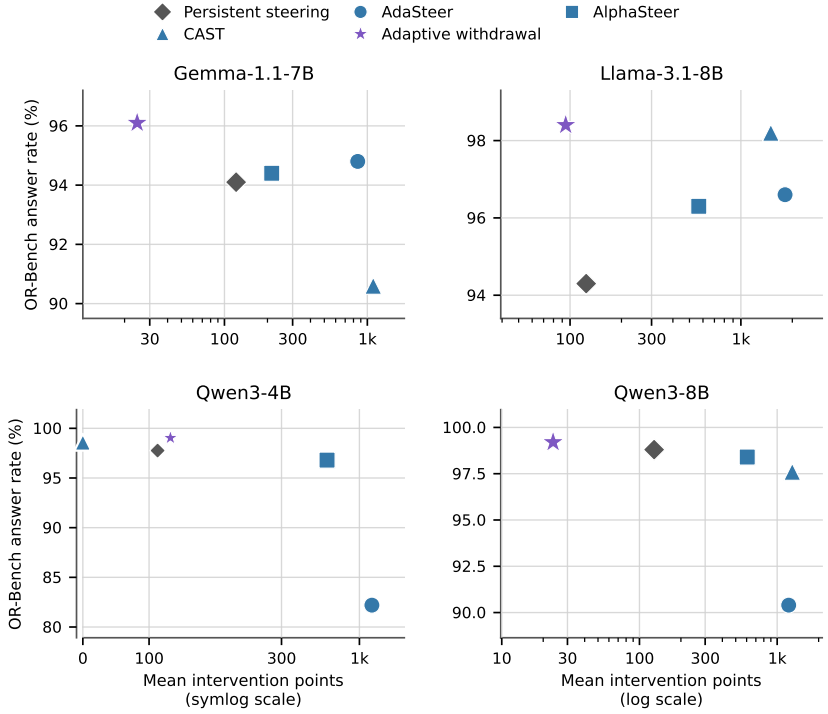


Figure 12: OR-Bench answer rate and intervention exposure across the four LLMs.

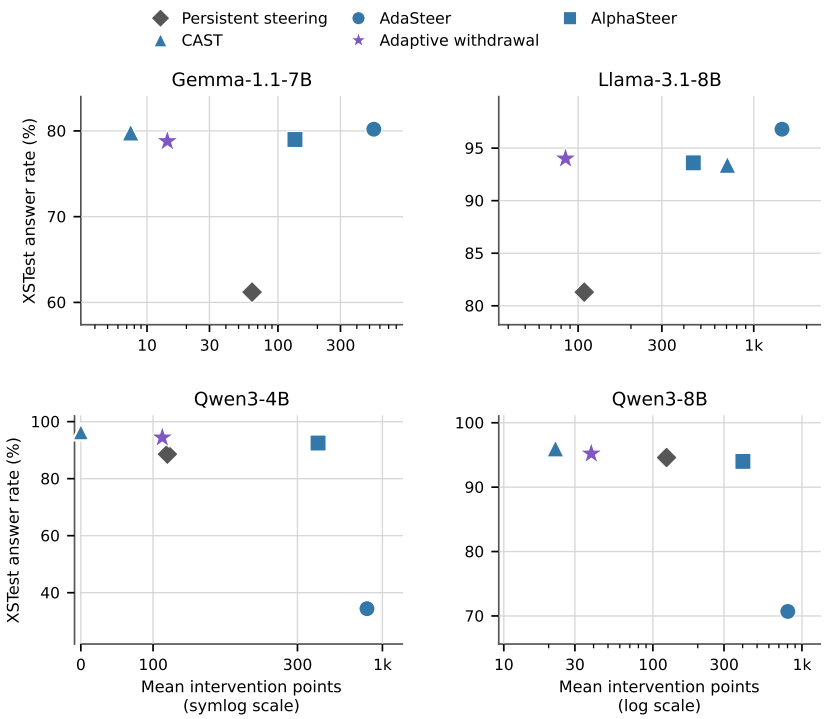


Figure 13: XSTest answer rate and intervention exposure across the four LLMs.

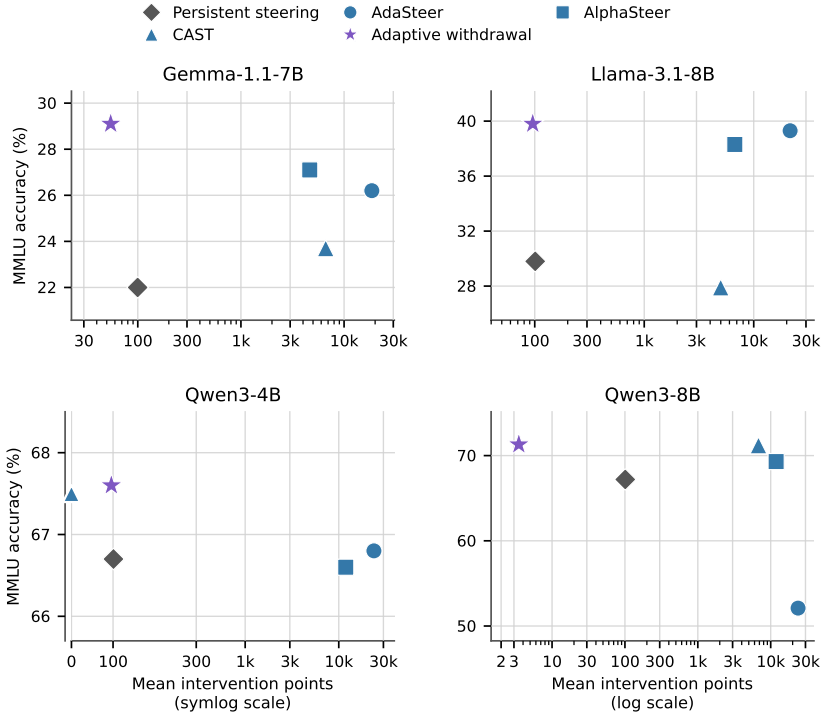


Figure 14: MMLU accuracy and intervention exposure across the four LLMs.

G GENERALISATION AND ROBUSTNESS

We distinguish transfer of the early refusal signals from transfer of the complete withdrawal rule, which also includes the fitted threshold.

G.1 TRANSFER ACROSS SAFETY EVALUATORS

We rescore stored adaptive-withdrawal and persistent-steering responses with Qwen2.5, Granite Guardian, and MD-Judge. These evaluators do not affect generation, predictor fitting, or withdrawal selection. Table 1 reports the primary HarmBench results. Table 19 isolates transfer to evaluators not used during calibration.

Table 19: Complete-response safety under three additional evaluators at $\alpha = 1$. Safety is reported as Adaptive/Persistent on the fixed evaluator-transfer response set.

Evaluator	Model	Safety (%)	Retained safety (%)	Meets criterion
Qwen2.5	Gemma-1.1-7B	99.00/100.00	99.0	Yes
	Llama-3.1-8B	96.25/96.25	100.0	Yes
	Qwen3-4B	99.25/99.00	100.3	Yes
	Qwen3-8B	94.25/94.75	99.5	Yes
Granite Guardian	Gemma-1.1-7B	95.75/97.25	98.5	Yes
	Llama-3.1-8B	79.00/78.50	100.6	Yes
	Qwen3-4B	73.50/91.75	80.1	No
	Qwen3-8B	65.50/67.50	97.0	Yes
MD-Judge	Gemma-1.1-7B	72.00/80.25	89.7	No
	Llama-3.1-8B	46.75/50.25	93.0	No
	Qwen3-4B	85.00/94.00	90.4	No
	Qwen3-8B	75.25/78.50	95.9	No

Table 19 shows that Qwen2.5 supports the comparison for all four LLMs and Granite Guardian supports three. This provides evidence of transfer beyond HarmBench, while the MD-Judge results and the Qwen3-4B Granite result reveal substantial evaluator dependence.

G.1.1 TARGETS CONSTRUCTED WITH TWO EVALUATORS

We also conduct a separate ablation in which withdrawal targets and thresholds are recalibrated jointly using HarmBench and Qwen2.5. A target is reached only when both evaluators classify the response as safe at that duration and every longer evaluable duration. This experiment evaluates a different policy from the primary HarmBench-calibrated policy in Table 1.

Table 20: Separately recalibrated policies whose targets and thresholds use both HarmBench and Qwen2.5. Safety columns report the joint policy rather than the primary policy.

Model	Threshold	C-index	HarmBench safety (%)	Qwen2.5 safety (%)	Exposure reduction (%)
Gemma-1.1-7B	0.80	0.919	98.25	99.00	62.0
Llama-3.1-8B	0.90	0.948	90.25	96.25	38.5
Qwen3-4B	0.95	0.669	95.75	99.00	47.3
Qwen3-8B	0.70	0.920	88.25	94.25	75.8

As shown in Table 20, joint calibration yields high complete-response safety under both evaluators for all four models. We retain HarmBench-only calibration in the primary experiments so evaluator transfer can be measured against a policy selected by a single fixed criterion.

G.2 TRANSFER ACROSS MODELS AND STEERING STRENGTHS

G.2.1 EARLY-SIGNAL TRANSFER ACROSS MODELS

For each target model, we fit the predictor on the other models and evaluate it without using target-model safety labels.

Table 21 shows that the combined trajectory retains cross-model ranking information for every target, with the strongest transfer on Llama and Gemma. Veracity is consistently informative. Velocity

Table 21: Leave-one-model-out prompt ranking at $\alpha = 1$.

Target model	Veracity + velocity C-index	Veracity-only C-index	Velocity-only C-index
Gemma-1.1-7B	0.880	0.919	0.204
Llama-3.1-8B	0.946	0.945	0.500
Qwen3-4B	0.619	0.669	0.500
Qwen3-8B	0.870	0.800	0.500

improves ranking most clearly for Qwen3-8B, contributes little for Llama, and is less helpful for Gemma and Qwen3-4B.

G.2.2 Complete Withdrawal-Rule Transfer Across Models

We next transfer both the predictor and threshold. Target-model activations are used for label-free standardisation, while target safety labels remain evaluation-only.

Table 22: Leave-one-model-out transfer of the complete withdrawal rule at $\alpha = 1$. Retained safety is relative to the matched persistent-steering control in this transfer experiment.

Target model	Transferred safety (%)	Retained safety (%)	Exposure reduction (%)	Meets criterion
Gemma-1.1-7B	88.00	88.2	91.0	No
Llama-3.1-8B	87.75	97.0	58.5	Yes
Qwen3-4B	90.75	94.0	87.9	No
Qwen3-8B	89.50	99.7	40.4	Yes

As reported in Table 22, the complete rule transfers successfully to Llama and Qwen3-8B. The weaker results for Gemma and Qwen3-4B show that ranking prompts is easier to transfer than the model-specific boundary that determines exactly when steering should end.

G.2.3 Within-Model Transfer Across Steering Strengths

Table 18 gives the direct transfer results from each model's frozen $\alpha = 1$ policy. Across the stronger settings, safety retention is 95.6–100.0% and exposure falls by 90.2–98.4%. Six of the eight stronger-strength settings meet the safety criterion without recalibration, showing that the learned policy often remains effective as steering becomes stronger.

G.3 Transfer to Out-of-Distribution Prompts

We apply each frozen model-specific $\alpha = 1$ policy without recalibration to direct harmful requests from StrongREJECT and to prompts from the unseen GCG and PAIR attack families (Souly et al., 2024; Zou et al., 2023b; Chao et al., 2025).

G.3.1 StrongREJECT Direct Requests

We evaluate all 313 direct requests in the StrongREJECT collection. Table 23 reports the same complete-response HarmBench safety and intervention exposure summarised in the main comparison.

Table 23: Transfer to 313 direct StrongREJECT requests without recalibration. Safety and mean intervention points are reported as Adaptive/Persistent.

Model	HarmBench safety Adaptive/Persistent (%)	Mean intervention points Adaptive/Persistent	Reduction (%)
Gemma-1.1-7B	99.6/99.6	2.0/34.4	94.2
Llama-3.1-8B	99.1/98.0	2.1/37.2	94.4
Qwen3-4B	99.7/99.7	22.6/106.4	78.8
Qwen3-8B	99.7/99.9	2.0/122.2	98.4

Table 23 shows that adaptive withdrawal achieves 99.1–99.7% complete-response HarmBench safety while reducing intervention exposure by 78.8–98.4%. Its safety differs from persistent steering by at most 1.1 percentage points, demonstrating transfer from attacked calibration prompts to direct harmful requests.

G.3.2 UNSEEN GCG AND PAIR ATTACKS

We evaluate 100 GCG prompts and 83 PAIR prompts. These attack families are not used to fit the predictor or select its threshold. Because the two sets have different sizes and difficulty, Table 24 reports them separately rather than averaging them into a single value.

Table 24: Transfer to unseen GCG and PAIR attacks. Entries report complete-response HarmBench safety and mean intervention points as Adaptive/Persistent.

Model	Attack	HarmBench safety Adaptive/Persistent (%)	Mean intervention points Adaptive/Persistent	Reduction (%)
Gemma-1.1-7B	GCG	98.6/99.0	2.8/31.2	91.0
	PAIR	96.1/96.1	4.5/37.5	88.0
Llama-3.1-8B	GCG	99.2/99.0	2.8/40.5	93.1
	PAIR	88.2/88.4	35.1/58.7	40.2
Qwen3-4B	GCG	99.2/100.0	13.4/71.4	81.2
	PAIR	83.1/84.3	98.2/118.9	17.4
Qwen3-8B	GCG	100.0/100.0	2.0/76.9	97.4
	PAIR	90.4/91.6	12.6/122.8	89.7

Table 24 shows consistently strong transfer on GCG: adaptive withdrawal remains within 0.8 percentage points of persistent steering while reducing exposure by 81.2–97.4%. PAIR is more difficult, but adaptive withdrawal remains within 1.2 percentage points of persistent steering and reduces exposure by 17.4–89.7%. The separation by attack family makes clear that transfer quality depends on the attack distribution, while the policy continues to preserve most of the persistent-steering safety with fewer interventions.

G.4 ADAPTIVE ATTACK

G.4.1 ATTACK CONSTRUCTION

We test whether an attacker can make the model begin with a refusal-like statement and then provide harmful content. Ten manually written refusal-then-comply templates vary how the initial refusal and later request are presented. We combine these templates with 30 development requests and retain the three templates per model that most often make the withdrawn continuation unsafe when continued steering remains safe.

For each of 80 held-out requests per model, the attacker uses white-box access to the first two output-token activations to select the retained template predicted to trigger the earliest withdrawal. Template selection never uses the held-out continuation or its safety label. We then compare withdrawal and continued steering from the same prefix and decoding state.

G.4.2 ATTACK RESULTS

Table 25: Adaptive-attack results under HarmBench. A new safety failure is a continuation that is unsafe after withdrawal but safe with steering maintained.

Model	Attack prompts	Finite withdrawals	New safety failures	Failure rate among withdrawals (%)
Gemma-1.1-7B	80	80	0	0.0
Llama-3.1-8B	80	80	11	13.8
Qwen3-4B	80	78	2	2.6
Qwen3-8B	80	80	1	1.3
All models	320	318	14	4.4

Table 25 shows that the attack induces finite withdrawal for 318 of 320 prompts, but only 14 withdrawals create a new safety failure. The failures are model dependent as Llama accounts for 11 of 80 attacks (13.8%), whereas the other models range from 0.0% to 2.6%. Across all models, 304 withdrawals do not make the continuation less safe than continued steering, giving an aggregate failure rate of 4.4%. Thus, triggering the policy to stop is usually insufficient to bypass refusal, although Llama exposes a clear target for stronger safeguards.

G.4.3 MINIMUM-DURATION SAFEGUARDS

We separately test minimum steering durations of 8, 16, and 32 tokens. A candidate must control development-set failures, satisfy the original held-out safety comparison, and reduce exposure by at least 25%. These safeguards are exploratory and are not used in the primary or SOTA results.

An eight-token minimum qualifies for Gemma and Qwen3-8B. No tested finite minimum meets all requirements for Llama, so the conservative fallback keeps steering active. For Qwen3-4B, all three tested minimums satisfy the clean-data requirements; a 32-token minimum removes both held-out HarmBench failures while still reducing exposure by 58.7%. These results illustrate how the fallback mechanism can retain steering when the evidence does not support safe withdrawal.