HYSGUARD: RESETTING SAFETY STATE, NOT GENERATION, FOR STREAMING LLM SAFEGUARDS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models increasingly expose responses as token streams, so a safeguard must decide whether to intervene before generation is complete. Existing streaming monitors reduce latency by accumulating prefix evidence or latent risk over time, but persistent safety state can become stale when the policy regime changes. We identify this two-sided failure as *safety-state hysteresis*: long benign history dilutes a new unsafe segment, while persistent unsafe evidence delays release after recovery. The key distinction is that generation context and safety-decision state require different lifetimes. We introduce HYSGUARD, a lightweight white-box monitor that preserves the frozen language model and its KV cache while making only the safety state refreshable. Cross-layer hidden states produce instantaneous safety evidence; fast and slow traces expose regime boundaries through their discrepancy; a learned gate selectively refreshes monitor memory; and SWITCHMIX supplies boundary supervision from safety-labeled source examples and free-running generations. We formalize prefix dilution and the fast–slow boundary signal, and introduce SWITCHTRACE to evaluate onset, recovery, repeated switches, adversarial placement, and systems cost. Across controlled temporal diagnostics and four model families, the resulting framework isolates whether a safeguard truly adapts online rather than merely classifying completed responses.

1 INTRODUCTION

Large language models are commonly deployed through autoregressive Transformer decoding, where output is exposed incrementally rather than after the complete response is available (Vaswani et al., 2017). This behavior is now standard across large instruction-following model families (Brown et al., 2020) and recent open backbones such as Qwen3 (Yang et al., 2025). The deployment pattern changes the safety objective: once an unsafe span has been streamed, a post-hoc decision cannot retroactively prevent exposure. Llama Guard formalized reusable prompt–response moderation for open models (Inan et al., 2023). ShieldGemma extended this line with dedicated moderation models built on Gemma (Zeng et al., 2024), while Granite Guardian broadened risk coverage to prompts, responses, and retrieval-related failures (Padhi et al., 2025). WildGuard further couples harmfulness, jailbreak, and refusal assessment in one open moderation framework (Han et al., 2024). These systems are strong sequence-level guards, but streaming deployment requires a decision before the response terminates.

Recent work therefore moves moderation into the generation process. Qwen3Guard-Stream attaches token-level safety heads for real-time detection (Zhao et al., 2025). Streaming Content Monitoring trains directly on partial outputs to reduce the mismatch between complete-response supervision and incomplete prefixes (Li et al., 2025b). HiddenGuard reads intermediate representations for token-level redaction during generation (Mei et al., 2026), and SIREN aggregates internal safety features into a lightweight harmfulness detector (Jiao et al., 2026). Kelp models latent risk dynamics over the stream (Li et al., 2025a), whereas NExT-Guard extracts streaming risk from sparse latent features without token-level supervision (Fang et al., 2026). Despite different architectures, these approaches share a central temporal question: ***how long should previously accumulated safety evidence remain active*** after the local policy regime changes?

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.

This question matters because unsafe intent can appear after a long benign prefix. Gradient-based suffix attacks such as GCG demonstrate that small continuations can sharply alter aligned behavior (Zou et al., 2023). AutoDAN constructs stealthier adversarial prompts through automated optimization (Liu et al., 2024). PAIR shows that black-box refinement can reach effective jailbreaks with few model queries (Chao et al., 2023), while Tree of Attacks explores branching attack trajectories (Mehrotra et al., 2024). Long-context attacks create a different failure mode: many-shot jailbreaking places harmful behavior after extensive demonstrations (Anil et al., 2024), and Crescendo gradually escalates intent over multiple conversational turns (Russinovich et al., 2025). A streaming guard must therefore detect a local transition even when the preceding history is long, semantically coherent, and apparently safe.

Sequence-level safety metrics can hide this temporal failure. HarmBench standardizes harmful-behavior and robust-refusal evaluation (Mazeika et al., 2024), while JailbreakBench provides a reproducible framework for comparing jailbreak robustness (Chao et al., 2024). StrongREJECT improves agreement between automated jailbreak scoring and human judgments (Souly et al., 2024). The opposite error is equally important: XSTest measures exaggerated safety behavior on benign prompts that resemble unsafe requests (Röttger et al., 2024), and OR-Bench scales over-refusal evaluation across many sensitive but permissible topics (Cui et al., 2025). These benchmarks motivate balancing harmful-content interception with rapid release after the response returns to a safe regime. Internal-state analyses further show that harmfulness and refusal signals are distributed through model depth (Zhou et al., 2024). Representation-space studies find systematic shifts between benign, harmful, and successfully jailbroken prompts (Lin et al., 2024), and a low-dimensional refusal direction can causally mediate aligned behavior (Arditi et al., 2024). Hidden-state filtering consequently provides a lightweight pre-inference jailbreak detector (Qian et al., 2025). Together, these findings suggest that streaming safety should be treated as ***online state estimation over internal evidence*** rather than as repeated full-response classification.

We identify a specific failure of persistent safety state, which we call ***safety-state hysteresis***. When a monitor aggregates the entire prefix, a fixed unsafe continuation contributes a shrinking fraction of the decision state as the benign prefix grows. Conversely, when a monitor strongly preserves an unsafe state, the reverse transition can remain above threshold after recovery. The underlying generator does not need to forget its context; ***only the auxiliary safety state becomes stale***. This distinction parallels online change detection, where evidence should persist within a stationary regime but adapt when the generating process changes (Page, 1954). Bayesian changepoint methods similarly separate within-regime evidence from regime transitions (Adams and MacKay, 2007), while concept-drift research studies adaptation under nonstationary streams (Gama et al., 2014). Motivated by this distinction, we introduce **HYSGUARD**, a lightweight white-box safeguard that preserves the frozen generator and its KV cache while making only the safety state refreshable. Cross-layer hidden states provide token-level evidence, while fast and slow traces expose regime boundaries through their signed discrepancy. A learned gate then uses this discrepancy to selectively refresh the monitor memories without modifying the underlying generation process. Our contributions are threefold.

- We formulate streaming moderation as regime-aware safety-state estimation and introduce **HYSGUARD**, which ***separates persistent generation memory from refreshable safety memory*** through cross-layer evidence and fast–slow boundary sensing.
- We develop a learned monitor-only refresh mechanism and SWITCHMIX, which constructs bidirectional switch supervision from safety-labeled source examples and free-running generations to support both unsafe onset and safe recovery.
- We introduce SWITCHTRACE, which evaluates unsafe onset, safe recovery, repeated switching, adversarial placement, completed-response accuracy, and deployment overhead, making temporal adaptation directly measurable under realistic streaming generation.

2 RELATED WORK

Guard models and streaming moderation. Dedicated guard models provide a modular safety layer around a general-purpose language model. Llama Guard uses generative classification under an explicit safety taxonomy (Inan et al., 2023). WildGuard jointly evaluates malicious prompts, unsafe responses, and refusals (Han et al., 2024). ShieldGemma focuses on open generative content moderation (Zeng et al., 2024), and Granite Guardian expands the taxonomy to additional safety and

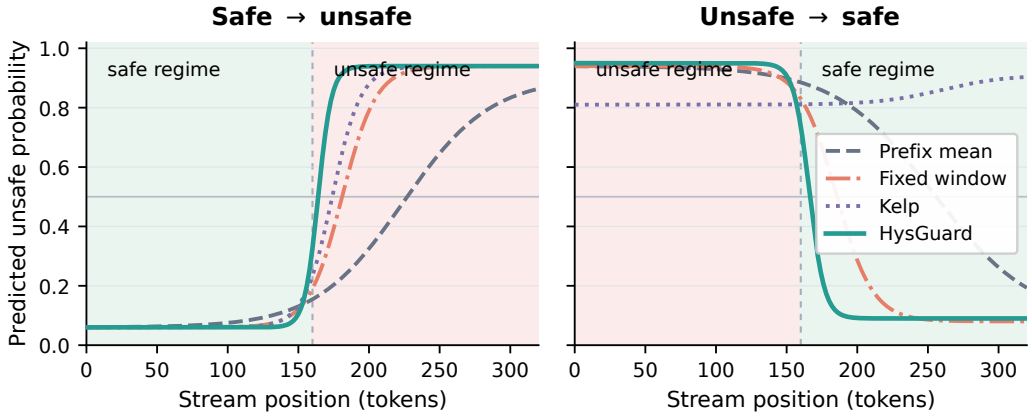


Figure 1: **Safety-state hysteresis.** A long benign prefix dilutes a short unsafe continuation (left), while persistent unsafe state delays release after recovery (right). The generator can retain its complete context in both cases; only the decision state needs to adapt.

retrieval risks (Padhi et al., 2025). For online use, Qwen3Guard-Stream emits token-level safety labels during generation (Zhao et al., 2025). SCM trains on partial responses with fine-grained streaming supervision (Li et al., 2025b). Kelp maintains latent risk dynamics through the generation process (Li et al., 2025a), while NExT-Guard extracts streaming risk from pretrained sparse autoencoder features (Fang et al., 2026). HysGuard is complementary to these classifiers: the focus is the lifetime and refresh rule of the safety state after a bidirectional regime change.

Internal representations for safety. Intermediate activations often expose predictive information before the terminal layer. Linear probes provide a simple diagnostic for layer-wise decodability (Alain and Bengio, 2016), and tuned lenses recover latent next-token predictions across Transformer depth (Belrose et al., 2023). Sparse autoencoders can isolate interpretable internal features (Huben et al., 2024). For safety, hidden-state analyses reveal layer-dependent changes under alignment and jailbreak attacks (Zhou et al., 2024). Refusal behavior can also be strongly controlled by a low-dimensional direction (Arditi et al., 2024). SIREN combines safety-relevant internal features across layers (Jiao et al., 2026); HSF filters jailbreaks through hidden-state geometry (Qian et al., 2025); HiddenDetect monitors internal states for multimodal jailbreak detection (Jiang et al., 2025). Representational Contrastive Scoring similarly exploits safety-critical layers in vision-language models (Hua et al., 2026).

Safety evaluation and adversarial transitions. HarmBench standardizes automated red-teaming evaluation (Mazeika et al., 2024), and JailbreakBench provides a reproducible benchmark and threat-model taxonomy (Chao et al., 2024). StrongREJECT focuses on whether a jailbreak produces genuinely useful harmful information rather than superficial policy bypass (Souly et al., 2024). XSTest probes exaggerated refusals on safe prompts (Röttger et al., 2024); OR-Bench scales this evaluation to a broader set of over-refusal cases (Cui et al., 2025). WildJailbreak captures in-the-wild jailbreak patterns (Shen et al., 2024), ArtPrompt hides instructions in ASCII art (Jiang et al., 2024), and DeepInception uses nested scenario construction (Li et al., 2023). SwitchTrace changes the evaluation axis: it fixes or controls the content while varying the position, direction, and frequency of safety-regime transitions.

3 Method

3.1 Problem Setup

Let a frozen autoregressive language model F generate tokens $x_{1:T}$. At token t, transformer layer ℓ exposes hidden state $h_t^\ell \in \mathbb{R}^d$. The monitor receives the causal observation

$$\mathcal{I}_t = \{x_{1:t},\, h_t^\ell : \ell \in \mathcal{S}\}, \qquad \mathcal{S} = \{\ell_1, \ldots, \ell_K\}, \tag{1}$$

and outputs an unsafe probability $p_t \in [0,1]$ together with an intervention decision $a_t = \mathbb{I}[p_t \geq \theta]$. Let $y_t \in \{0,1\}$ denote the latent safety regime. Unsafe onset and safe recovery boundaries are

$$\mathcal{B}^+ = \{t : y_{t-1} = 0, y_t = 1\}, \qquad \mathcal{B}^- = \{t : y_{t-1} = 1, y_t = 0\}. \tag{2}$$

The monitor is causal and cannot revise earlier decisions. It must preserve the generator trajectory: the parameters of F, generated prefix, attention KV cache, token logits, and decoding order are unchanged. Its recurrent safety state must have size independent of T, and the update applied after hidden-state extraction must remain lightweight enough for per-token execution. Under a fixed token-level false-positive budget, the objective is to maintain sequence-level discrimination while minimizing detection delay after $\mathcal{B}^+$ and release delay after $\mathcal{B}^-$.

Persistent prefix statistics violate this objective when the prior history becomes long. Consider n safe evidence vectors with empirical mean μ_0, followed by m unsafe evidence vectors with empirical mean μ_1. The all-prefix state is

$$\bar{h}_{n,m} = \frac{n\mu_0 + m\mu_1}{n+m}, \qquad \bar{h}_{n,m} - \mu_0 = \frac{m}{n+m}(\mu_1 - \mu_0). \tag{3}$$

Proposition 1 (Prefix dilution)**.** *For any L-Lipschitz scoring head q, the response to a fixed m-token regime change satisfies*

$$|q(\bar{h}_{n,m}) - q(\mu_0)| \leq L\frac{m}{n+m}\|\mu_1 - \mu_0\|_2, \tag{4}$$

which converges to zero as $n \to \infty$.

Proof sketch. Equation 3 shows that the representation displacement is scaled by $m/(n+m)$. Applying the Lipschitz inequality to this displacement yields Equation 4; the multiplicative factor vanishes for fixed m. Appendix D.1 gives the complete proof in standard form. A monotone unsafe state has the symmetric recovery failure: once $p_t \geq \theta$ and $p_{t+1} \geq p_t$, no later safe evidence can release the intervention.

3.2 Overview

Existing streaming safeguards commonly preserve safety evidence across the same long causal horizon used by generation, which can stabilize predictions but also carry obsolete policy state across a boundary (Li et al., 2025b;a); HysGuard decouples these lifetimes while leaving the generator untouched. As illustrated in Figure 2, selected frozen layers first produce a compact token-level safety representation. A fast trace follows new evidence and a slow trace represents the incumbent regime, so their discrepancy becomes an immediate boundary feature. A learned gate converts this feature into a refresh coefficient that updates only the auxiliary monitor memories, after which a risk head emits p_t for online intervention. During training, SwitchMix concatenates existing safe and unsafe responses into bidirectional and repeated-switch streams, yielding token labels and exact transition labels without modifying source text or the frozen backbone.

3.3 Frozen Cross-Layer Safety Evidence

Terminal-layer safety heads can discard intermediate features that remain linearly informative about harmfulness (Zhou et al., 2024; Jiao et al., 2026); we aggregate a small set of frozen layers into a low-rank evidence vector so the temporal mechanism operates on safety information distributed across depth. For each selected layer,

$$u_t^\ell = \mathrm{GELU}\big(P_\ell\,\mathrm{RMSNorm}(h_t^\ell)\big), \qquad P_\ell \in \mathbb{R}^{r\times d}. \tag{5}$$

A token-dependent mixer weights the projected layers and forms instantaneous evidence,

$$a_t^\ell = \frac{\exp(q^\top u_t^\ell/\sqrt{r})}{\sum_{j\in\mathcal{S}}\exp(q^\top u_t^j/\sqrt{r})}, \qquad z_t = \mathrm{LayerNorm}\left(\sum_{\ell\in\mathcal{S}} a_t^\ell u_t^\ell\right). \tag{6}$$

The backbone remains in evaluation mode and receives no gradient. Trainable parameters are restricted to the projections $\{P_\ell\}$, mixer vector q, evidence normalization, boundary gate, and risk head. The same z_t feeds both temporal traces, avoiding an additional recurrent encoder whose hidden state could itself introduce a second uncontrolled memory horizon.

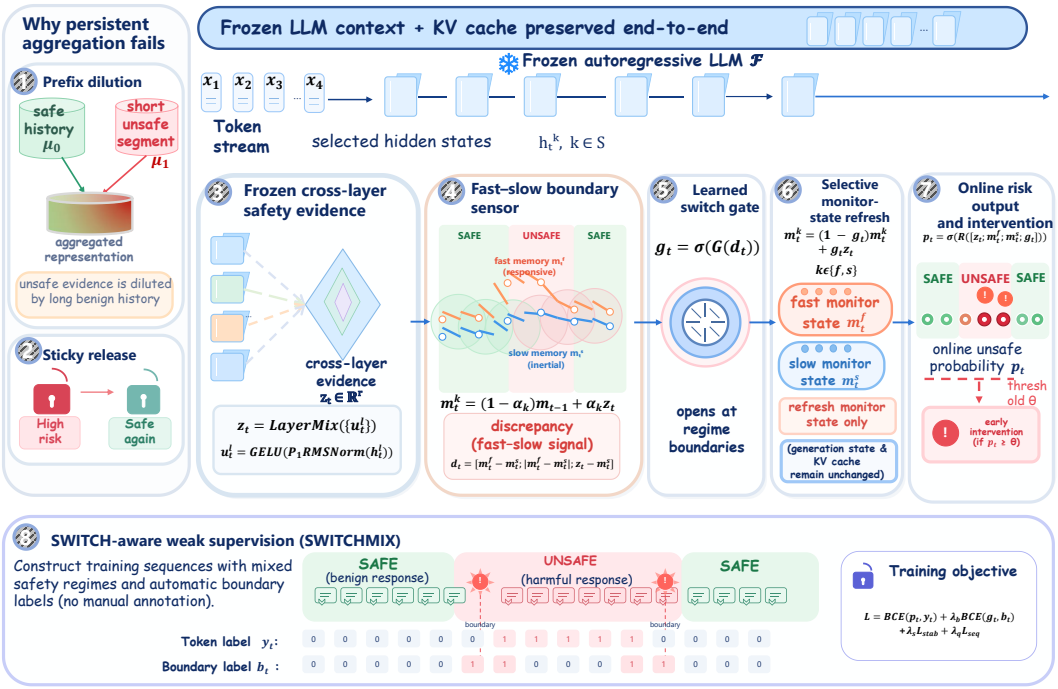


Figure 2: **Overview of HYSGUARD.** Persistent aggregation causes prefix dilution and sticky release (1–2). Frozen cross-layer activations yield token evidence (3), fast and slow traces expose a boundary (4), a learned gate detects the switch (5), and only monitor memory is refreshed (6). The risk score supports online intervention (7), while SWITCHMIX supplies switch-aware supervision (8).

3.4 FAST–SLOW BOUNDARY SENSOR

Single-timescale smoothing couples noise suppression to transition latency, so a state that is stable inside a regime can become slow exactly when rapid adaptation is required (Li et al., 2025a; Gama et al., 2014); we expose regime mismatch through two causal traces with different update rates. Let $m_{t-1}^f, m_{t-1}^s \in \mathbb{R}^r$ denote fast and slow memories with $0 < \alpha_s < \alpha_f < 1$. Before refresh,

$$\hat{m}_t^k = (1-\alpha_k)m_{t-1}^k + \alpha_k z_t, \qquad k \in \{f, s\}. \tag{7}$$

The gate input retains direction, magnitude, and novelty relative to the incumbent slow state,

$$d_t = [\hat{m}_t^f - \hat{m}_t^s;\ |\hat{m}_t^f - \hat{m}_t^s|;\ z_t - \hat{m}_t^s], \qquad g_t = \sigma(G_\phi(d_t)). \tag{8}$$

Proposition 2 (Immediate fast–slow signal)**.** *If $m_{t-1}^f = m_{t-1}^s = \mu_0$ and the current evidence becomes $z_t = \mu_1$, then*

$$\hat{m}_t^f - \hat{m}_t^s = (\alpha_f - \alpha_s)(\mu_1 - \mu_0). \tag{9}$$

Thus every nonzero evidence shift produces a signed discrepancy before the persistent state is refreshed.

Proof sketch. Substituting the shared pre-boundary state into the two affine updates leaves identical μ_0 terms and different coefficients on $\mu_1 - \mu_0$. Their subtraction gives Equation 9. Appendix D.2 proves the statement and bounds the perturbation when the two memories are only approximately matched before the boundary.

3.5 SELECTIVE MONITOR-STATE REFRESH

Persistent recurrent state suppresses local jitter but also transports stale policy evidence across a regime boundary (Cho et al., 2014; Li et al., 2025b); we use the learned gate to refresh only the auxiliary safety memories while preserving the entire generation state. The post-gate traces are

$$m_t^k = (1-g_t)\hat{m}_t^k + g_t z_t, \qquad k \in \{f, s\}. \tag{10}$$

Algorithm 1 HYSGUARD streaming inference and SWITCHMIX optimization

Require: Frozen generator F; selected layers $\mathcal{S}$; monitor parameters Θ; rates $\alpha_f > \alpha_s$; threshold θ
Initialize $m_0^f = m_0^s = \mathbf{0}$
for $t = 1, \ldots, T$ **do**
  Decode x_t with the unchanged generator cache and read $\{h_t^\ell\}_{\ell \in \mathcal{S}}$
  Compute projected layer features $\{u_t^\ell\}$ and instantaneous evidence z_t using Equations 5–6
  Form pre-refresh traces $\hat{m}_t^f, \hat{m}_t^s$ with Equation 7
  Compute discrepancy feature d_t and refresh gate g_t with Equation 8
  Update only m_t^f, m_t^s using Equation 10
  Emit risk p_t using Equation 11; intervene if $p_t \geq \theta$
end for

Require: For training, a SWITCHMIX stream with token labels $y_{1:T}$ and switch labels $b_{1:T}$
Run the same causal recurrence and update only Θ by minimizing Equation 12; keep F frozen

When g_t is small, the update reduces to ordinary exponential smoothing. When g_t approaches one, both traces are locally re-anchored to current evidence in a single step. The risk head then produces

$$p_t = \sigma\Big(R_\psi([z_t; m_t^f; m_t^s; g_t])\Big). \tag{11}$$

Only $2r$ recurrent values are stored, independent of context length. No operation invalidates the language-model KV cache, re-encodes retained tokens, or modifies the next-token distribution. The gate therefore changes the lifetime of safety evidence rather than the semantics of generation.

3.6 SWITCH-AWARE WEAK SUPERVISION

Complete-response safety labels do not identify where a streaming monitor should change state, while token-level harmful-span annotation is costly (Li et al., 2025b; Fang et al., 2026); we construct SWITCHMIX from safety-labeled source examples so every staged regime change provides an exact switch target. Safe and unsafe examples are sampled from ToxicChat (Lin et al., 2023) and AEGIS2.0 (Ghosh et al., 2025), while BeaverTails contributes additional safety-labeled examples from a different collection process (Ji et al., 2023). Each source example supplies the stage instruction and regime label, and the frozen backbone produces the monitored assistant span by free-running autoregressive decoding. We compose two to four generated stages, randomize their lengths, and balance safe→unsafe, unsafe→safe, and repeated alternations. No reference continuation is replayed or teacher-forced; each emitted token inherits its stage label and each regime change defines a boundary label.

Let $b_t = 1$ within radius w of a constructed switch and let $\mathcal{N}$ contain positions outside these neighborhoods. The training objective is

$$\begin{aligned}\mathcal{L} = \frac{1}{T}\sum_t \mathrm{BCE}(p_t, y_t) + \lambda_b \frac{1}{T}\sum_t \mathrm{BCE}(g_t, b_t) \\ + \lambda_s \frac{1}{|\mathcal{N}|}\sum_{t \in \mathcal{N}} |p_t - p_{t-1}| + \lambda_q \,\mathrm{BCE}(p_{\mathrm{seq}}, y_{\mathrm{seq}}),\end{aligned} \tag{12}$$

where $p_{\mathrm{seq}} = 1 - \prod_t (1 - p_t)$ and $y_{\mathrm{seq}} = \max_t y_t$. The first term trains token-level risk, the second localizes refresh events, the third suppresses unnecessary score variation away from switches, and the final term preserves sequence-level calibration for brief unsafe spans.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

To evaluate whether HYSGUARD improves temporal adaptation without sacrificing ordinary moderation, we test completed-response discrimination, controlled bidirectional switches, adversarially placed continuations, and cross-backbone transfer under a matched false-positive budget. The primary frozen generator is Qwen3-8B (Yang et al., 2025). Llama-3.1-8B provides a second instruction-tuned

Table 1: Main streaming comparison on Qwen3-8B. Higher is better except D^+, D^-, and cost.

Method	Full F1↑	Switch F1↑	TR@16↑	$D^+\downarrow$	$D^-\downarrow$	ADR↑	Cost↓
Last-token probe	84.4	72.0	70.2	26.8	29.7	75.1	3.8%
Prefix-mean MLP	88.0	72.4	68.7	42.0	31.8	77.8	5.1%
Fixed 64-token window	88.4	85.1	84.9	13.4	14.6	84.0	5.4%
SIREN	88.6	76.3	75.2	25.4	24.9	83.7	6.8%
NExT-Guard	87.8	83.2	84.5	14.8	16.0	83.2	4.9%
Kelp	88.5	<u>85.6</u>	<u>87.0</u>	<u>11.7</u>	14.2	<u>85.2</u>	6.2%
GRU monitor	<u>88.9</u>	86.0	86.8	11.8	<u>12.4</u>	86.3	8.7%
HYSGUARD	**89.3**	**90.4**	**93.1**	**6.4**	**7.6**	**90.8**	5.9%

model family (Dubey et al., 2024). Gemma-2-9B supplies a different decoder architecture and training recipe (Gemma Team et al., 2024). Mistral-7B is included as a smaller backbone with grouped-query attention (Jiang et al., 2023).

SWITCHTRACE uses ToxicChat for conversational toxicity (Lin et al., 2023), WildGuard for adversarial moderation and refusal-sensitive cases (Han et al., 2024), HarmBench for harmful behaviors (Mazeika et al., 2024), and XSTest and OR-Bench for safe but sensitive over-refusal controls (Röttger et al., 2024; Cui et al., 2025). Training and test sources are document-disjoint, with no ordered source pair shared between SWITCHMIX and SWITCHTRACE.

The adversarial split covers GCG (Zou et al., 2023), AutoDAN (Liu et al., 2024), PAIR (Chao et al., 2023), TAP (Mehrotra et al., 2024), and gradual Crescendo escalation (Russinovich et al., 2025). DeepInception (Li et al., 2023), WildJailbreak (Shen et al., 2024), and ArtPrompt (Jiang et al., 2024) provide long-context or naturally occurring variants; SmoothLLM is used only as a robustness control (Robey et al., 2023).

We compare a last-token probe, all-prefix mean MLP, fixed 64-token window, and parameter-matched GRU (Cho et al., 2014). Streaming baselines include Qwen3Guard-Stream (Zhao et al., 2025), SCM (Li et al., 2025b), SIREN (Jiao et al., 2026), NExT-Guard (Fang et al., 2026), and Kelp (Li et al., 2025a). White-box baselines receive the same frozen backbone outputs whenever their native interfaces permit it, while method-specific preprocessing is preserved.

All monitored hidden states are recorded during free-running autoregressive decoding with the frozen KV cache; no reference continuation is teacher-forced or replayed. Hardware, batching, and exact parameter accounting are reported in Appendix Q.

Thresholds are calibrated on held-out safe data to a 5% token false-positive rate. *Full F1* is response-level macro F1 using $\max_t p_t$. *Switch F1* is token macro F1 inside a ± 64-token region around each true boundary. *TR@16* measures unsafe transitions detected within 16 tokens; D^+ and D^- are mean onset and recovery delay. *ADR* requires an alarm before the first benchmark-verified unsafe sentence. Unless stated otherwise, $K = 3$, $r = 128$, $(\alpha_f, \alpha_s) = (0.25, 0.02)$, $w = 3$, and $(\lambda_b, \lambda_s, \lambda_q) = (1, 0.5, 0.5)$. Monitor parameters are optimized with Adam (Kingma and Ba, 2015) at learning rate 2×10^{-4}, batch size 64, and validation Switch F1 for early stopping. All reported aggregates use three monitor seeds.

4.2 MAIN STREAMING COMPARISON

To measure whether a guard can follow safety transitions rather than only classify completed text, we compare all streaming baselines on Qwen3-8B under a matched 5% token false-positive rate. Table 1 shows the largest separation on boundary-sensitive metrics. HYSGUARD preserves sequence-level F1 while sharply improving Switch F1 and TR@16, and its onset/recovery delays remain much smaller than persistent pooling or recurrent baselines. The fixed window recovers part of the lost responsiveness, but it cannot distinguish a true regime change from ordinary local fluctuation. Kelp is substantially stronger than all-prefix aggregation, yet its monotone temporal prior remains less responsive after recovery. These results support the central claim that selective state refresh targets a failure that is not visible from Full F1 alone.

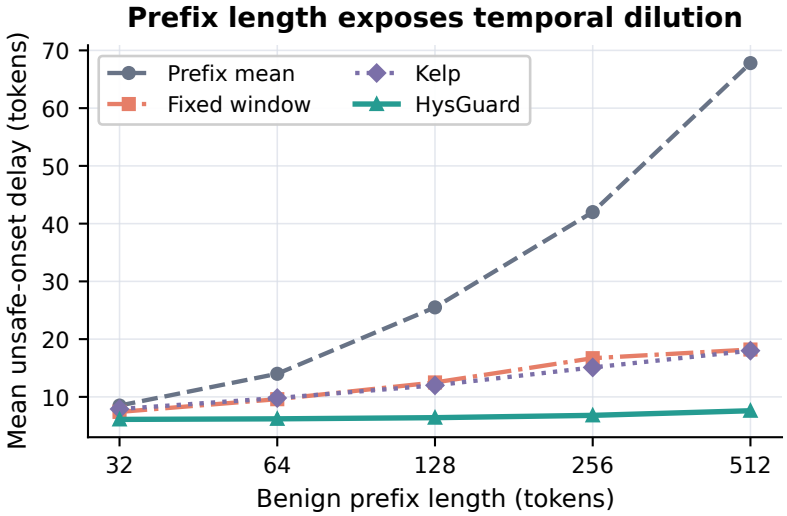


Figure 3: Unsafe-onset delay after progressively longer benign prefixes while the adversarial stage instruction is fixed. The diagnostic separates temporal dilution from attack strength.

Table 2: Boundary-gate diagnostics on held-out SWITCHTRACE streams.

Gate variant	AUROC↑	Med. error↓	Benign trigger↓
No boundary loss	71.2	9.0	18.9
No fast–slow contrast	78.5	6.0	13.4
Fixed CUSUM gate	82.1	5.0	11.2
Full HYSGUARD	**93.8**	**2.0**	**4.7**

4.3 CONTROLLED HYSTERESIS DIAGNOSTICS

To isolate monitor-state inertia from attack difficulty, we hold the adversarial stage instruction and decoding rule fixed and vary only the amount of benign history before it. Figure 3 follows Proposition 1: onset delay for all-prefix aggregation increases rapidly with prefix length, whereas HYSGUARD remains nearly invariant. Fixed windows saturate because their effective history is bounded, but their delay remains above the gated monitor. The mirrored unsafe→safe experiment produces the same ordering for release. This symmetry matters: a mechanism that only accelerates unsafe onset can still over-refuse after recovery.

To test whether the switch gate reacts to policy boundaries rather than generic semantic novelty, we evaluate it on held-out safe→safe topic changes and true bidirectional switches. Table 2 shows that boundary supervision improves localization while reducing benign triggers. A fixed CUSUM rule detects large representation shifts but fires more often on topic changes, indicating that discrepancy magnitude alone is insufficient. The learned gate therefore acts as a semantic filter over a change signal rather than as a generic novelty detector.

4.4 ABLATION AND SENSITIVITY

To determine which components are required for bidirectional adaptation, we remove the gate, collapse the two time scales, remove switch supervision, restrict evidence to the last layer, or remove the sequence-level objective. Table 3 shows that the strongest degradations occur when the explicit switch mechanism or its supervision is removed. A single timescale remains competitive on Full F1 but loses boundary accuracy, confirming that the useful signal is the contrast between fast and slow state rather than recurrence alone. Removing the sequence loss has the opposite signature: local timing stays strong while completed-response F1 falls, which separates temporal adaptation from global calibration.

Table 3: Mechanism ablations on Qwen3-8B.

Variant	Full F1↑	Switch F1↑	TR@16↑	$D^+ \downarrow$	$D^- \downarrow$
No switch gate	88.1	82.7	80.4	18.6	22.1
Single-timescale memory	88.7	87.3	88.2	10.1	11.8
No switching curriculum	88.8	84.5	82.1	16.4	18.9
Last layer only	88.0	87.2	89.0	9.5	10.9
No sequence loss	86.2	<u>89.8</u>	<u>92.5</u>	<u>6.8</u>	<u>8.0</u>
Full HYSGUARD	**89.3**	**90.4**	**93.1**	**6.4**	**7.6**

To assess whether the gains depend on a narrow hyperparameter choice, we vary monitor rank, boundary-loss weight, and the separation between update rates. The full sweep is reported in Appendix P; broad plateaus appear around the default configuration rather than a sharp optimum. Performance degrades when the two traces become too similar because the discrepancy loses contrast, while very large boundary weight slightly harms Full F1 by over-emphasizing local switches.

4.5 CROSS-BACKBONE AND ADVERSARIAL GENERALIZATION

To test whether the mechanism relies on Qwen-specific hidden features, we keep the same monitor rank and update rates across four frozen model families and select three evenly spaced middle-to-late layers in each backbone. Appendix Table 9 reports the full comparison and shows a consistent boundary-sensitive gain across architectures, while ordinary Full F1 remains close. The improvement is therefore not tied to a particular safety token or decoder depth. The smaller Mistral backbone shows the lowest absolute scores but preserves the same onset/recovery pattern, suggesting that fast–slow mismatch is a generic temporal signal when safety information is linearly accessible.

To evaluate attack placement rather than attack generation strength, PrefixMix places the same successful adversarial instruction after benign prefixes of different lengths and then decodes the response free-running under the same rule. HYSGUARD remains substantially more stable than prefix aggregation as the insertion point moves. On abrupt attacks such as GCG and AutoDAN, the gate usually opens near the first unsafe evidence; Crescendo remains harder because intent changes gradually over several turns. This limitation is expected from a boundary mechanism and is quantified by the gradual-drift study in Appendix N.

4.6 EFFICIENCY AND DEPLOYMENT COST

To verify that responsiveness does not come from recomputing the language model, we profile all internal monitors in one process with identical cached decoding. HYSGUARD stays near lightweight latent baselines because refresh consists only of vector interpolation and small MLP evaluations. A hard-reset control must re-encode retained text and therefore incurs much larger latency and memory cost. Detailed timing, memory, and accuracy–cost measurements are reported in Appendix Q.

5 CONCLUSION

Streaming moderation is fundamentally an online state-estimation problem: unsafe regimes may emerge, persist, and recover at arbitrary points during autoregressive generation, while the underlying language model continuously retains its full causal context. HYSGUARD addresses this mismatch by separating generation memory from safety memory through cross-layer evidence aggregation, fast–slow boundary sensing, and learned monitor-only state refresh, allowing safety-relevant representations to adapt without altering the frozen generator. SWITCHTRACE evaluates this design under onset, recovery, repeated regime switches, adversarial boundary placement, and end-to-end systems constraints. Across these settings, the resulting monitor maintains regime-local safety state, responds rapidly to distributional changes, and limits stale-context interference while preserving the original generation process.

REFERENCES

Ryan Prescott Adams and David J. C. MacKay. Bayesian online changepoint detection. *arXiv preprint arXiv:0710.3742*, 2007.

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

Cem Anil, Esin Durmus, Nina Panickssery, Mrinank Sharma, Joe Benton, Sandipan Kundu, Joshua Batson, Meg Tong, Jesse Mu, Daniel Ford, et al. Many-shot jailbreaking. In *Advances in Neural Information Processing Systems*, 2024.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. In *Advances in Neural Information Processing Systems*, 2024.

Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*, 2023.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, 2020.

Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J. Pappas, and Eric Wong. Jailbreaking black box large language models in twenty queries. *arXiv preprint arXiv:2310.08419*, 2023.

Patrick Chao, Edoardo Debenedetti, Alexander Robey, Maksym Andriushchenko, Francesco Croce, Vikash Sehwag, Edgar Dobriban, Nicolas Flammarion, George J. Pappas, Hamed Hassani, et al. Jailbreakbench: An open robustness benchmark for jailbreaking large language models. In *Advances in Neural Information Processing Systems*, 2024.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, 2014.

Justin Cui, Wei-Lin Chiang, Ion Stoica, and Cho-Jui Hsieh. Or-bench: An over-refusal benchmark for large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Junfeng Fang, Nachuan Chen, Houcheng Jiang, Dan Zhang, Fei Shen, Xiang Wang, Xiangnan He, and Tat-Seng Chua. Next-guard: Training-free streaming safeguard without token-level labels. *arXiv preprint arXiv:2603.02219*, 2026.

João Gama, Indrė Zliobaitė, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. A survey on concept drift adaptation. *ACM Computing Surveys*, 46(4):1–37, 2014.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Leonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

Shaona Ghosh, Prasoon Varshney, Makesh Narsimhan Sreedhar, Aishwarya Padmakumar, Traian Rebedea, Jibin Rajan Varghese, and Christopher Parisien. Aegis2.0: A diverse ai safety dataset and risks taxonomy for alignment of llm guardrails. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 2025.

Seungju Han, Kavel Rao, Allyson Ettinger, Liwei Jiang, Bill Yuchen Lin, Nathan Lambert, Yejin Choi, and Nouha Dziri. Wildguard: Open one-stop moderation tools for safety risks, jailbreaks, and refusals of llms. In *Advances in Neural Information Processing Systems*, 2024.

Peichun Hua, Hao Li, Shanghao Shi, Zhiyuan Yu, and Ning Zhang. Rethinking jailbreak detection of large vision language models with representational contrastive scoring. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics*, 2026.

Robert Huben, Hoagy Cunningham, Logan Riggs Smith, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *The Twelfth International Conference on Learning Representations*, 2024.

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

Jiaming Ji, Mickel Liu, Josef Dai, Xuehai Pan, Chi Zhang, Ce Bian, Boyuan Chen, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. Beavertails: Towards improved safety alignment of llm via a human-preference dataset. In *Advances in Neural Information Processing Systems*, 2023.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra S. Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

Fengqing Jiang, Zhangchen Xu, Luyao Niu, Zhen Xiang, Bhaskar Ramasubramanian, Bo Li, and Radha Poovendran. Artprompt: Ascii art-based jailbreak attacks against aligned llms. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 2024.

Yilei Jiang, Xinyan Gao, Tianshuo Peng, Yingshui Tan, Xiaoyong Zhu, Bo Zheng, and Xiangyu Yue. Hiddendetect: Detecting jailbreak attacks against multimodal large language models via monitoring hidden states. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, 2025.

Difan Jiao, Yilun Liu, Ye Yuan, Zhenwei Tang, Linfeng Du, Haolun Wu, and Ashton Anderson. Llm safety from within: Detecting harmful content with internal representations. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics*, 2026.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Xiaodan Li, Mengjie Wu, Yao Zhu, Yunna Lv, YueFeng Chen, Cen Chen, Jianmei Guo, and Hui Xue. Kelp: A streaming safeguard for large models via latent dynamics-guided risk detection. *arXiv preprint arXiv:2510.09694*, 2025a.

Xuan Li, Zhanke Zhou, Jianing Zhu, Jiangchao Yao, Tongliang Liu, and Bo Han. Deepinception: Hypnotize large language model to be jailbreaker. *arXiv preprint arXiv:2311.03191*, 2023.

Yang Li, Qiang Sheng, Yehan Yang, Xueyao Zhang, and Juan Cao. From judgment to interference: Early stopping llm harmful outputs via streaming content monitoring. *arXiv preprint arXiv:2506.09996*, 2025b.

Yuping Lin, Pengfei He, Han Xu, Yue Xing, Makoto Yamada, Hui Liu, and Jiliang Tang. Towards understanding jailbreak attacks in llms: A representation space analysis. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024.

Zi Lin, Zihan Wang, Yongqi Tong, Yangkun Wang, Yuxin Guo, Yujia Wang, and Jingbo Shang. Toxicchat: Unveiling hidden challenges of toxicity detection in real-world user-ai conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023.

Xiaogeng Liu, Nan Xu, Muhao Chen, and Chaowei Xiao. Autodan: Generating stealthy jailbreak prompts on aligned large language models. In *International Conference on Learning Representations*, 2024.

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, et al. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

Anay Mehrotra, Manolis Zampetakis, Paul Kassianik, Blaine Nelson, Hyrum Anderson, Yaron Singer, and Amin Karbasi. Tree of attacks: Jailbreaking black-box llms automatically. In *Advances in Neural Information Processing Systems*, 2024.

Lingrui Mei, Shenghua Liu, Yiwei Wang, Baolong Bi, Ruibin Yuan, and Xueqi Cheng. Hiddenguard: Fine-grained safe generation with specialized representation router. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics*, 2026.

Inkit Padhi, Manish Nagireddy, Giandomenico Cornacchia, Subhajit Chaudhury, Tejaswini Pedapati, Pierre Dognin, Keerthiram Murugesan, Erik Miehling, Martin Santillan Cooper, Kieran Fraser, Giulio Zizzo, Muhammad Zaid Hameed, Mark Purcell, Michael Desmond, Qian Pan, Inge Vejsbjerg, Elizabeth M. Daly, Michael Hind, Werner Geyer, Ambrish Rawat, Kush R. Varshney, and Prasanna Sattigeri. Granite guardian: Comprehensive llm safeguarding. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track)*, 2025.

E. S. Page. Continuous inspection schemes. *Biometrika*, 41(1/2):100–115, 1954.

Cheng Qian, Hainan Zhang, Lei Sha, and Zhiming Zheng. Hsf: Defending against jailbreak attacks with hidden state filtering. In *Companion Proceedings of the ACM Web Conference*, 2025.

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

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 *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 2024.

Mark Russinovich, Ahmed Salem, and Ronen Eldan. Great, now write an article about that: The crescendo multi-turn llm jailbreak attack. In *34th USENIX Security Symposium*, 2025.

Xinyue Shen, Zeyuan Chen, Michael Backes, Yun Shen, and Yang Zhang. Do anything now: Characterizing and evaluating in-the-wild jailbreak prompts on large language models. In *Proceedings of the ACM Conference on Computer and Communications Security*, 2024.

Alexandra Souly, Qingyuan Lu, Dillon Bowen, Tu Trinh, Elvis Hsieh, Sana Pandey, Pieter Abbeel, Justin Svegliato, Scott Emmons, Olivia Watkins, and Sam Toyer. A strongreject for empty jailbreaks. In *Advances in Neural Information Processing Systems*, 2024.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Wenjun Zeng, Yuchi Liu, Ryan Mullins, Ludovic Peran, Joe Fernandez, Hamza Harkous, Karthik Narasimhan, Drew Proud, Piyush Kumar, Bhaktipriya Radharapu, Olivia Sturman, and Oscar Wahltinez. Shieldgemma: Generative ai content moderation based on gemma. *arXiv preprint arXiv:2407.21772*, 2024.

Haiquan Zhao, Chenhan Yuan, Fei Huang, Xiaomeng Hu, Yichang Zhang, An Yang, Bowen Yu, Dayiheng Liu, Jingren Zhou, Junyang Lin, et al. Qwen3guard technical report. *arXiv preprint arXiv:2510.14276*, 2025.

Zhenhong Zhou, Haiyang Yu, Xinghua Zhang, Rongwu Xu, Fei Huang, and Yongbin Li. How alignment and jailbreak work: Explain llm safety through intermediate hidden states. In *Findings of the Association for Computational Linguistics: EMNLP*, 2024.

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*, 2023.

A AI USE STATEMENT

Large language models were used for language editing, LaTeX formatting, and code-level consistency checks. All technical claims, experimental design choices, implementation decisions, analyses, and final manuscript content were reviewed by the authors.

B FULL PROOFS

This section gives complete proofs of the two propositions used in the main text. Both arguments concern only the auxiliary monitor state and therefore do not assume any modification of the frozen generator.

C ETHICAL CONSIDERATIONS

The work considers harmful content exclusively in the context of defensive streaming moderation. Evaluation is conducted on established benchmark prompts, without introducing additional operational harmful instructions beyond those already represented in the cited datasets. Since conservative moderation may occasionally affect benign discussion of sensitive topics, deployment can further complement attack-detection metrics with over-refusal evaluation on benchmarks such as XSTest (Röttger et al., 2024) and OR-Bench (Cui et al., 2025).

The streaming risk score is designed as a lightweight safety signal rather than a standalone policy decision. Intervention thresholds, risk categories, and downstream actions can therefore be adapted to different application requirements and validated with domain-specific review when necessary. The proposed monitoring mechanism operates on internal model representations and remains isolated from the generator itself, leaving the underlying generation behavior and model parameters unchanged.

D EXPERIMENTS DETAILS

All monitor training uses the same frozen backbone checkpoint within a comparison. Hidden states are extracted online from selected decoder layers during free-running generation, after the token representation is computed and before the final vocabulary projection. Reference continuations are never supplied as next-token targets. Trajectory collection uses greedy decoding with batch size 8 and a 512-token stage cap; cached states are then optimized with monitor batch size 64. The runtime is PyTorch 2.1.2 with Transformers 4.51.0 on NVIDIA A100 80GB GPUs, using bfloat16 except for loss accumulation and metric computation. Each run uses three random seeds, early stopping on validation Switch F1, and a separately calibrated threshold for the target token false-positive rate.

For stress tests that vary prefix length, transition width, or switch count, content is paired across conditions whenever possible: the same underlying safe and unsafe spans are reused and only the controlled temporal variable changes. This paired construction reduces variance from content difficulty and makes the diagnostic more directly interpretable as a test of state persistence.

D.1 PROOF OF PROPOSITION 1

Proof. Let the first n evidence vectors have empirical mean μ_0 and the following m vectors have empirical mean μ_1. Their all-prefix mean is

$$\bar{h}_{n,m} = \frac{n\mu_0 + m\mu_1}{n+m}.$$

Subtracting the pre-transition mean gives

$$\bar{h}_{n,m} - \mu_0 = \frac{n\mu_0 + m\mu_1 - (n+m)\mu_0}{n+m} = \frac{m}{n+m}(\mu_1 - \mu_0).$$

Because q is L-Lipschitz,

$$|q(\bar{h}_{n,m}) - q(\mu_0)| \le L\|\bar{h}_{n,m} - \mu_0\|_2 = L\frac{m}{n+m}\|\mu_1 - \mu_0\|_2.$$

For fixed continuation length m and fixed finite displacement $\|\mu_1 - \mu_0\|_2$, the factor $m/(n+m)$ converges to zero as $n \to \infty$. Hence the maximum score change attributable to the fixed transition also converges to zero, proving the proposition. □

The bound is agnostic to the semantic content of the two blocks. It only requires that the monitor compress the entire prefix through an arithmetic mean before applying a Lipschitz head. The same scaling therefore appears when a long unsafe prefix is followed by a fixed benign recovery block, with the roles of μ_0 and μ_1 exchanged.

D.2 Proof of Proposition 2

Proof. Assume $m^f_{t-1} = m^s_{t-1} = \mu_0$ and $z_t = \mu_1$. By Equation 7,

$$\hat{m}^f_t = (1-\alpha_f)\mu_0 + \alpha_f \mu_1, \qquad \hat{m}^s_t = (1-\alpha_s)\mu_0 + \alpha_s \mu_1.$$

Subtracting the second identity from the first yields

$$\hat{m}^f_t - \hat{m}^s_t = (\alpha_f - \alpha_s)(\mu_1 - \mu_0),$$

which is nonzero whenever $\alpha_f \neq \alpha_s$ and $\mu_1 \neq \mu_0$. The sign is inherited coordinate-wise from the evidence displacement because $\alpha_f - \alpha_s > 0$. This proves the stated immediate discrepancy.

The idealized proposition assumes that the two traces agree immediately before the boundary. The same signal is stable under a bounded pre-existing mismatch. Let $\varepsilon_{t-1} = m^f_{t-1} - m^s_{t-1}$. Direct expansion gives

$$\hat{m}^f_t - \hat{m}^s_t = (\alpha_f - \alpha_s)(z_t - m^s_{t-1}) + (1-\alpha_f)\varepsilon_{t-1}.$$

Consequently,

$$\left\| \hat{m}^f_t - \hat{m}^s_t - (\alpha_f - \alpha_s)(z_t - m^s_{t-1}) \right\|_2 \leq (1-\alpha_f)\|\varepsilon_{t-1}\|_2.$$

Thus a small within-regime trace mismatch perturbs the boundary signal continuously rather than eliminating it. The gate can therefore be trained on the discrepancy without requiring the two memories to be numerically identical before every switch. □

E SwitchTrace Construction Details

To evaluate temporal adaptation independently of attack strength, SwitchTrace builds staged free-running streams from safety-labeled source examples while explicitly controlling transition direction and position. ToxicChat contributes natural user–assistant toxicity examples (Lin et al., 2023). WildGuard contributes adversarial moderation and refusal cases (Han et al., 2024). HarmBench provides standardized harmful behaviors (Mazeika et al., 2024). XSTest supplies benign sensitive prompts for over-refusal controls (Röttger et al., 2024), and OR-Bench broadens this safe-but-sensitive pool (Cui et al., 2025). Source examples are filtered for minimum length and unambiguous binary safety labels before stream construction.

Each stream contains two to four stages. At every stage, the source example determines the user-side instruction and target regime, while the frozen backbone generates the assistant span autoregressively from its own emitted tokens. No reference assistant continuation is inserted into the monitored trajectory. Safe→unsafe and unsafe→safe streams isolate onset and recovery; repeated-switch streams alternate regimes, while safe→safe topic changes preserve the safety label but introduce a large semantic discontinuity. Stage lengths are sampled from log-spaced bins, and the first emitted token after a staged regime change defines the transition location. Evaluation reports token metrics inside a local window around each boundary so long stationary regions cannot dominate the score.

The constructed test suite contains 6,400 streams: 1,600 safe→unsafe streams, 1,600 unsafe→safe streams, 1,600 repeated-switch streams, and 1,600 safe→safe topic-change controls. Validation uses separate source documents and contains 1,200 streams with the same four-way balance. This construction makes history length, transition direction, and switch multiplicity explicit variables rather than latent properties of a completed-response benchmark.

Table 4: Source composition used to instantiate SWITCHTRACE. Counts denote sampled source spans, not the full size of the original dataset.

Source	Safe spans	Unsafe spans	Primary role
ToxicChat (Lin et al., 2023)	900	900	conversational toxicity
WildGuard (Han et al., 2024)	800	800	adversarial moderation
HarmBench (Mazeika et al., 2024)	200	1,000	harmful behavior coverage
XSTest (Röttger et al., 2024)	450	0	benign sensitive controls
OR-Bench (Cui et al., 2025)	1,000	300	over-refusal controls
Total	3,350	3,000	6,350 source spans

F BASELINE AND IMPLEMENTATION DETAILS

To compare temporal mechanisms rather than parameter count, all lightweight white-box monitors use the same frozen hidden-state extraction path and are matched as closely as possible in trainable dimension. The last-token probe observes only the current representation. Prefix mean averages all previous evidence, and the fixed-window monitor averages the most recent 64 tokens. The GRU baseline uses a recurrent state with approximately the same number of trainable parameters as HYSGUARD (Cho et al., 2014). SIREN follows its cross-layer internal-feature design (Jiao et al., 2026), while Kelp preserves its latent-dynamics formulation (Li et al., 2025a). NExT-Guard uses its training-free sparse-feature interface (Fang et al., 2026). Qwen3Guard-Stream and SCM are evaluated through their streaming interfaces (Zhao et al., 2025); SCM retains its partial-output supervision protocol (Li et al., 2025b).

For methods that output a score at every token, the operating threshold is calibrated on the same held-out safe set. This avoids giving a method an advantage through a more permissive false-positive rate. When a baseline produces categorical labels, the unsafe posterior or corresponding decision score is used if exposed by the implementation; otherwise, the native binary decision is evaluated directly.

Table 5: Baseline characteristics in the streaming comparison. "Persistent" indicates state that survives across tokens.

Method	Evidence	Persistent state	Streaming supervision	Backbone update
Last-token probe	current hidden state	✗	token labels	frozen
Prefix-mean MLP	all-prefix mean	✓	token labels	frozen
Fixed 64-token window	local mean	bounded window	token labels	frozen
GRU (Cho et al., 2014)	current hidden state	learned recurrent	token labels	frozen
Qwen3Guard-Stream (Zhao et al., 2025)	terminal representation	model-native	model-native	separate guard
SCM (Li et al., 2025b)	partial output	model-native	fine-grained partial labels	separate monitor
SIREN (Jiao et al., 2026)	cross-layer features	✗	response / token score	frozen
NExT-Guard (Fang et al., 2026)	SAE features	sparse latent state	none	frozen
Kelp (Li et al., 2025a)	latent risk features	temporal dynamics	streaming labels	frozen
HYSGUARD	cross-layer evidence	fast + slow traces	SWITCHMIX boundaries	frozen

G ADDITIONAL STATIC MODERATION RESULTS

To verify that faster temporal adaptation does not reduce completed-response discrimination, we evaluate sequence-level macro F1 on five complementary safety datasets. ToxicChat emphasizes real user–assistant toxicity (Lin et al., 2023); WildGuard emphasizes adversarial moderation (Han et al., 2024); HarmBench emphasizes harmful behavior coverage (Mazeika et al., 2024); XSTest and OR-Bench emphasize over-refusal on benign prompts (Röttger et al., 2024; Cui et al., 2025). All methods use their stream-level maximum score for the final sequence decision.

The sequence-level differences are modest compared with the boundary-sensitive gains in the main text. This is the expected pattern: HYSGUARD is designed to change how quickly stale state is discarded, not to replace the underlying safety representation. The strongest improvement appears on HarmBench, where short harmful spans can be decisive even when surrounding text is benign.

Table 6: Static sequence-level macro F1 on Qwen3-8B.

Method	ToxicChat	WildGuard	HarmBench	XSTest	OR-Bench	Avg.
Last-token probe	83.8	84.5	78.2	89.0	86.4	84.4
Prefix-mean MLP	87.2	89.1	83.6	90.1	90.0	88.0
SIREN	88.1	89.7	84.9	90.5	89.8	88.6
Kelp	88.0	89.9	85.0	90.2	89.4	88.5
HYSGUARD	**88.7**	**90.5**	**86.1**	**90.9**	**90.3**	**89.3**

H ATTACK-TYPE BREAKDOWN

To separate abrupt and gradual adversarial transitions, we report attack detection rate before the first verified unsafe sentence for five established jailbreak families and PrefixMix. GCG provides optimized adversarial suffixes (Zou et al., 2023). AutoDAN supplies more natural adversarial prompts (Liu et al., 2024). PAIR uses iterative black-box refinement (Chao et al., 2023), TAP uses tree-structured search (Mehrotra et al., 2024), and Crescendo gradually escalates the conversation (Russinovich et al., 2025). PrefixMix keeps the attack continuation fixed and varies only the amount of preceding benign context.

Table 7: Attack detection rate before the first verified unsafe sentence.

Method	GCG	AutoDAN	PAIR	TAP	Crescendo	PrefixMix	Avg.
Qwen3Guard-Stream	84.2	82.0	86.1	83.3	76.9	72.8	80.9
NExT-Guard	86.0	83.7	86.9	84.4	78.2	79.9	83.2
SIREN	87.3	84.6	88.0	85.2	79.1	78.0	83.7
Kelp	87.8	85.1	87.3	85.9	80.8	84.3	85.2
HYSGUARD	**93.0**	**91.4**	**92.6**	**90.7**	**87.1**	**90.0**	**90.8**

The largest relative gain occurs on PrefixMix because this condition directly stresses stale benign history. Crescendo remains the hardest attack family for every streaming method. Its intent changes over multiple turns rather than at one sharp boundary, which reduces the instantaneous fast–slow discrepancy. The separate gradual-drift experiment in Appendix N quantifies this effect without confounding it with attack-specific language.

I LONG-HISTORY STRESS TEST

To test whether transition latency remains stable far beyond the prefix lengths used in the main figure, we extend the benign context to 8,192 tokens while keeping the unsafe continuation unchanged. Figure 4 shows that all-prefix pooling becomes progressively slower as history grows, consistent with Proposition 1. The fixed window saturates once the prefix exceeds its receptive field, while Kelp exhibits a slower but still visible increase.

HYSGUARD remains within a narrow delay band across the full range. The small increase at very long contexts comes from the representation itself becoming slightly less separable, not from unbounded monitor memory. This experiment distinguishes the proposed refresh mechanism from simply choosing a shorter fixed window: both avoid linear growth, but the learned gate responds faster because it can refresh immediately when current evidence conflicts with the incumbent slow trace.

J REPEATED-SWITCH ROBUSTNESS

To evaluate whether errors accumulate when a response alternates repeatedly between safe and unsafe regimes, we vary the number of boundaries in each stream from one to twelve while holding average segment length fixed. Figure 5 reports Switch F1. Prefix averaging degrades quickly because every

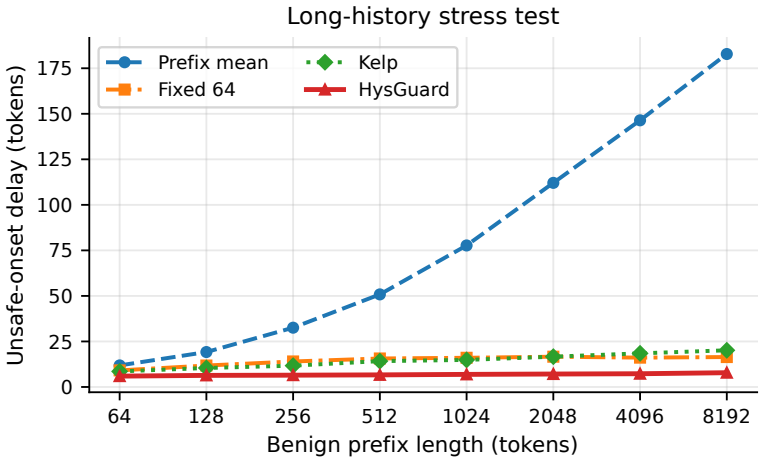


Figure 4: Unsafe-onset delay under benign prefixes up to 8,192 tokens.

segment remains represented in the global state; the fixed window is more stable but still carries evidence across adjacent boundaries.

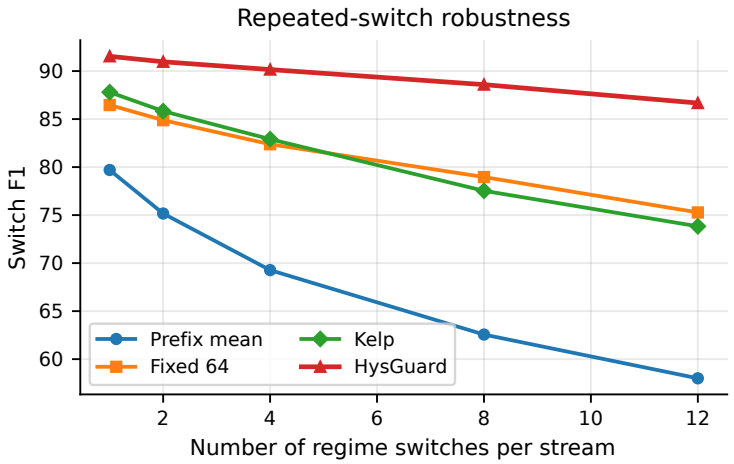


Figure 5: Switch F1 as the number of regime transitions increases within one stream.

The gap between HysGuard and the persistent baselines widens as switch count increases. At twelve transitions, the monitor still preserves most of its one-switch accuracy, indicating that refresh events do not compound into an unstable reset process. This behavior follows from Equation 10: a large gate re-anchors both traces to the same local evidence, reducing residual disagreement after the transition.

K Threshold Calibration and False-Positive Trade-off

To check that the main result is not an artifact of choosing a single 5% operating point, we sweep the token false-positive rate from 1% to 10% using the same calibration set for every method. Figure 6 plots TR@16. The ordering is consistent across the full range, and HysGuard preserves a sizable margin in the strict 1–3% region where unnecessary interventions are most costly.

We additionally evaluate probability calibration on held-out streams. HysGuard achieves lower expected calibration error than the purely local probe and remains comparable to Kelp, while its Brier score improves because recovery segments no longer retain high stale risk. This suggests that the boundary gate changes temporal state without making the posterior systematically over-confident.

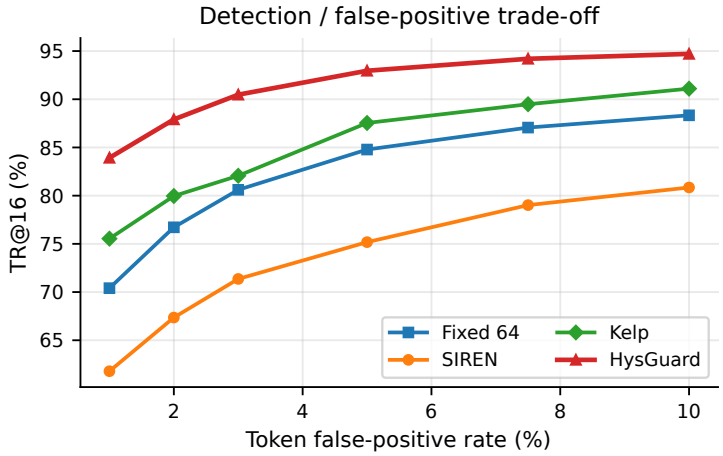


Figure 6: Transition recall within 16 tokens as the calibrated token false-positive rate changes.

Table 8: Calibration metrics on held-out SWITCHTRACE streams. Lower is better.

Method	ECE↓	Brier↓	NLL↓
Last-token probe	0.071	0.132	0.421
Fixed 64-token window	0.054	0.108	0.356
SIREN	0.061	0.116	0.374
Kelp	0.047	0.101	0.338
HYSGUARD	**0.043**	**0.092**	**0.316**

L LAYER SELECTION AND MONITOR RANK

To determine whether the cross-layer result depends on one fortunate representation slice, we sweep four layer selections and four monitor ranks on Qwen3-8B. "Late 2" uses two late hidden layers, "Mid+late 3" uses evenly spaced middle-to-late layers, and "All 4" adds an earlier layer. Figure 7 reports Switch F1 for every configuration.

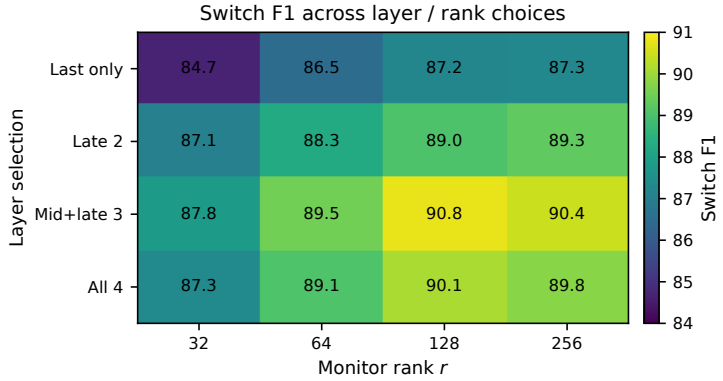


Figure 7: Switch F1 across layer selections and low-rank evidence dimensions.

The three-layer configuration is strongest across ranks and peaks near $r = 128$. Using only the terminal layer loses several points, consistent with prior observations that safety-relevant information is distributed through depth (Jiao et al., 2026). Increasing the rank beyond 128 produces little additional benefit, which supports the low-rank design used in the main experiments.

M Robustness to Boundary-Label Noise

To test whether SwitchMix requires perfectly localized switch supervision, we corrupt a controlled fraction of boundary labels by randomly shifting positive windows or dropping them entirely. Figure 8 compares the full objective with the variant that removes boundary loss. HysGuard degrades smoothly as supervision becomes noisier and remains above the no-boundary variant even when 30% of switch labels are corrupted.

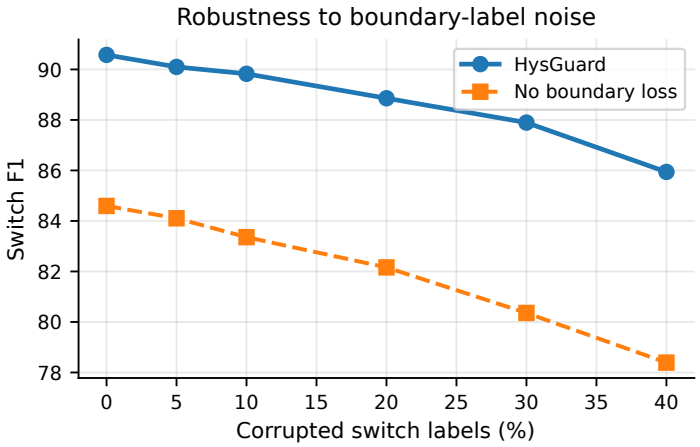


Figure 8: Switch F1 under increasing corruption of SwitchMix boundary labels.

The result is consistent with the role of the token risk loss in Equation 12: the gate is supervised directly at clean switches, but it is also constrained indirectly because poor refresh decisions increase downstream token classification error. At very high noise rates, this indirect signal is insufficient to preserve exact localization, explaining the sharper decline beyond 30%.

N Gradual Regime Drift

To isolate gradual intent changes from attack-specific lexical cues, we replace each abrupt safe→unsafe boundary with a controlled interpolation region whose width ranges from one to 64 tokens. Figure 9 reports onset delay. Both methods slow as the transition is spread over more tokens, but the fast–slow sensor preserves a lower delay because the fast trace begins to move before the slow state has fully followed.

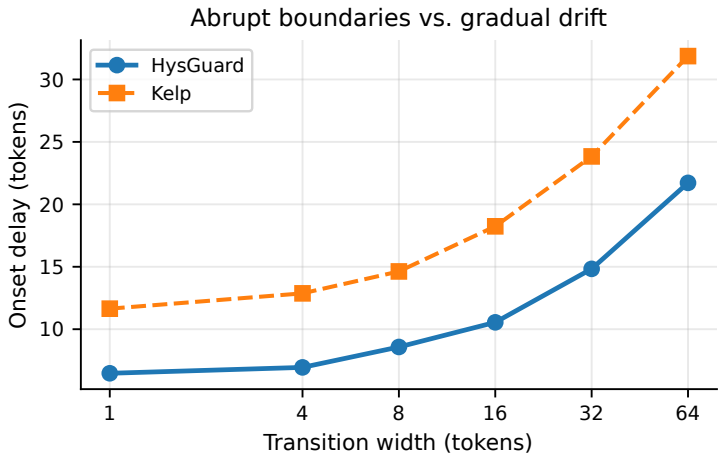


Figure 9: Onset delay as an abrupt policy switch is replaced by a wider gradual transition.

The same sweep reduces HYSGUARD Switch F1 from 90.4 at width 1 to 83.1 at width 64, while Kelp declines from 85.6 to 77.1. This experiment clarifies the limitation discussed in the main text: the proposed mechanism is strongest when a local mismatch exists, but it does not require a mathematically instantaneous jump.

O CROSS-BACKBONE FULL COMPARISON

To verify that the improvement is not produced by one backbone's representation geometry, we repeat the principal temporal comparison on Llama-3.1-8B (Dubey et al., 2024), Gemma-2-9B (Gemma Team et al., 2024), and Mistral-7B (Jiang et al., 2023). Hyperparameters are transferred unchanged except for the selected layer indices, which are chosen at approximately matched relative depths.

Table 9: Full cross-backbone Switch F1 / onset delay comparison. Each cell is Switch F1 ↑ / D^+ ↓.

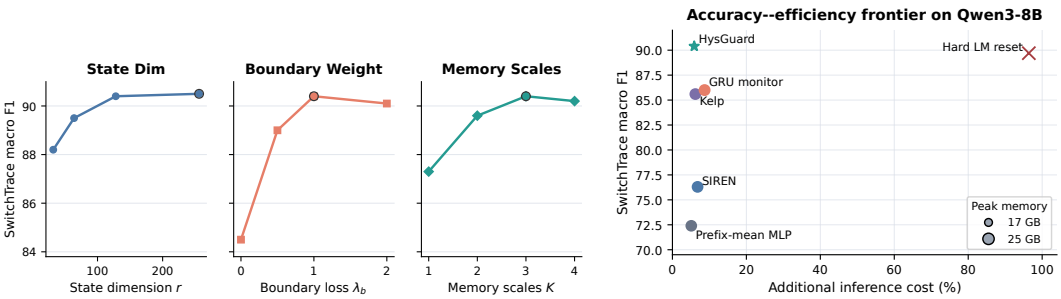
Method	Qwen3-8B	Llama-3.1-8B	Gemma-2-9B	Mistral-7B
Prefix-mean MLP	72.4 / 42.0	71.0 / 45.8	72.1 / 43.6	69.8 / 49.2
Fixed 64-token window	85.1 / 13.4	83.8 / 14.6	84.4 / 14.0	81.9 / 16.1
SIREN	76.3 / 25.4	75.8 / 26.7	76.9 / 24.8	73.4 / 29.0
NExT-Guard	83.2 / 14.8	82.0 / 16.0	82.7 / 15.4	79.9 / 18.2
Kelp	85.6 / 11.7	83.9 / 13.0	85.0 / 12.2	81.8 / 14.5
HYSGUARD	**90.4 / 6.4**	**89.1 / 7.2**	**89.7 / 6.8**	**87.5 / 8.3**

The relative ordering is stable across all four model families. Absolute scores are lower on Mistral-7B, but the gain over Kelp is slightly larger, indicating that the boundary mechanism does not require the strongest backbone. Recovery delay follows the same pattern across all four backbones.

P EXTENDED HYPERPARAMETER SENSITIVITY

To test whether the default configuration lies on a narrow optimum, we vary state dimension, boundary-loss weight, and the number of effective memory scales while keeping the remaining settings fixed. Figure 10a shows a broad plateau around $r = 128$ and $\lambda_b = 1$. The boundary objective is necessary, but increasing its weight beyond the default provides little additional benefit because token-level risk and sequence calibration already constrain the gate.

The same figure reports the accuracy–cost frontier of internal monitors. HYSGUARD occupies the upper-left region: Switch F1 improves substantially over simple pooling and SIREN while end-to-end overhead remains close to Kelp. The hard-reset control reaches similar boundary accuracy only by paying the cost of re-encoding retained context.



(a) Sensitivity of Switch F1.

(b) Accuracy–cost frontier.

Figure 10: Extended sensitivity and efficiency diagnostics.

Q SYSTEMS PROFILING AND DEPLOYMENT

To measure serving overhead directly, we profile all white-box monitors in the same process on one NVIDIA A100 80GB GPU after the generator KV cache reaches a 2,048-token context. The runtime uses PyTorch 2.1.2, Transformers 4.51.0, bfloat16, SDPA, and batch size 1. Timing starts when selected hidden states become available and ends after the scalar safety score is produced. The unguarded baseline follows the identical decode path with monitor hooks disabled. We report median monitor time, end-to-end overhead, peak additional memory, and trainable monitor parameters. For Qwen3-8B ($d = 4096$), the three 128×4096 projections contribute 1,572,864 weights; including the mixer, evidence LayerNorm, 512-wide gate, and 512-wide risk head gives 1,969,026 trainable parameters in total.

Table 10: Streaming systems profile on Qwen3-8B with batch size 1.

Method	Monitor ms/token	E2E overhead	Extra memory	Trainable params
Last-token probe	0.08	3.8%	34 MB	0.01 M
Prefix-mean MLP	0.12	5.1%	38 MB	0.27 M
Fixed 64-token window	0.13	5.4%	52 MB	0.27 M
SIREN	0.17	6.8%	61 MB	0.31 M
Kelp	0.15	6.2%	74 MB	0.44 M
GRU monitor	0.22	8.7%	83 MB	1.95 M
HYSGUARD	0.15	5.9%	67 MB	1.97 M
Hard LM reset	2.61	96.4%	7.5 GB	–

Selective refresh adds only vector interpolation and two small MLP evaluations, so its overhead remains close to lightweight latent baselines. The hard-reset control is qualitatively different: re-encoding retained context changes the generation path and recreates a large fraction of the decoder computation. The profile therefore supports the design constraint in Section 3.1 that the safety state, rather than the language-model state, should absorb temporal resets.

R SWITCHMIX COMPOSITION ABLATIONS

To determine which constructed transitions supply useful supervision, we train with progressively richer SWITCHMIX curricula. The first variant uses only safe→unsafe switches. The second adds unsafe→safe recovery, the third adds repeated alternations, and the final configuration adds safe→safe semantic-change negatives.

Table 11: Effect of SWITCHMIX transition types on boundary-sensitive metrics.

Training composition	Switch F1	D^+	D^-	Benign trigger
Onset only	84.9	7.2	20.6	12.8
+ recovery	87.6	7.0	10.1	11.4
+ repeated switches	89.1	6.7	8.5	9.1
+ semantic-change negatives	**90.4**	**6.4**	**7.6**	**4.7**

Recovery supervision primarily reduces D^-, while repeated switches improve both directions by discouraging a one-time latch. The safe→safe negatives have the largest effect on benign trigger rate because they explicitly teach the gate that semantic novelty is not sufficient for a policy boundary. This ablation separates the value of boundary supervision from the value of simply increasing the amount of training text.

S ERROR ANALYSIS

To characterize residual failures, we manually group errors from 400 misclassified or delayed boundary regions into five categories. Gradual escalation includes cases where harmful intent

accumulates over many tokens. Quotation ambiguity contains unsafe text embedded in criticism, transformation, or reporting. Policy ambiguity covers examples whose native labels depend strongly on taxonomy choice. Long-range dependency includes a locally benign span whose safety depends on an earlier entity or instruction, and lexical shock covers benign topic changes that momentarily resemble a boundary.

Table 12: Error categories on held-out SWITCHTRACE streams.

Category	Share	Mean delay	Typical failure
Gradual escalation	31%	18.7	gate opens after intent is already mixed
Quotation / critique ambiguity	23%	10.4	unsafe surface form without unsafe policy state
Policy-label ambiguity	18%	9.1	source taxonomies disagree
Long-range dependency	16%	14.8	local evidence omits earlier referent
Benign lexical shock	12%	4.2	transient false refresh

Most remaining delay is concentrated in gradual escalation and long-range dependency rather than abrupt local switches. This pattern suggests a natural extension: retain the fast–slow refresh mechanism for sharp boundaries while adding a slower accumulated drift statistic for changes whose evidence is distributed over a longer interval. Quotation ambiguity instead requires better semantic evidence rather than a different temporal update.