CANARY-CONTROLLED FEEDBACK INTERLEAVING FOR EMERGENT MISALIGNMENT IN LLM FINE-TUNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Fine-tuning large language models (LLMs) on narrowly defined tasks can lead to emergent misalignment, where undesirable behaviors appear on unrelated inputs even when the training data itself is benign. A common mitigation strategy is to interleave auxiliary safe data during training; however, existing approaches typically adopt fixed mixing schedules, which are unable to respond to the evolving nature of misalignment. We introduce a canary-controlled feedback interleaving approach that adjusts data mixing during training based on online risk signals. Specifically, a set of canary prompts is used to probe model behavior, producing a running estimate of misalignment risk via exponential smoothing. This signal drives a simple decision policy with hysteresis, allowing the training process to increase or relax intervention in a stable manner as conditions change. Evaluations on the Security EM benchmark with Qwen2.5-7B-Instruct show that the proposed approach reduces the misalignment rate from 7.15% to 5.39% (a 25% relative decrease) compared to fixed-ratio interleaving, while also achieving 4× lower variance across random seeds. Additional studies reveal that performance gains are primarily attributable to adaptive intervention timing, rather than the overall proportion of safe data, as fixed-schedule variants perform 36% worse under the same data budget. These findings indicate that effective alignment during fine-tuning depends less on static data composition and more on responsive adjustment during training, suggesting a shift toward adaptive strategies for improving robustness.

1 INTRODUCTION

Large language models (LLMs) are typically aligned through extensive post-training procedures to promote helpful, harmless, and honest behavior. Despite these efforts, recent studies have identified a counterintuitive failure mode: fine-tuning on narrowly scoped and seemingly benign tasks can trigger emergent misalignment, where undesirable behaviors surface on inputs unrelated to the training objective. This phenomenon is particularly concerning because it arises without explicitly harmful data, making it difficult to anticipate and prevent.

To address this issue, prior work has explored various mitigation strategies during fine-tuning. A widely adopted approach is to interleave auxiliary safe data with task-specific training samples, aiming to preserve alignment properties. Other methods focus on constraining parameter updates or preserving safety-relevant representations. While these techniques can reduce misalignment to some extent, they generally rely on fixed intervention strategies, applying the same level of correction throughout training. Such static designs overlook an important characteristic of emergent misalignment: its temporal variability.

Existing evidence suggests that misalignment does not develop gradually, but instead may appear abruptly at certain stages of training, resembling phase-transition-like behavior. This implies that the risk of misalignment is non-uniform over time, and therefore unlikely to be effectively handled by a constant intervention strength. A fixed strategy may underreact during high-risk periods or introduce unnecessary overhead when the model remains stable.

Motivated by this observation, we explore a dynamic intervention paradigm that adapts to the model's evolving state. We introduce a canary-guided feedback interleaving approach, which uses a small set of diagnostic prompts to continuously probe model behavior during training. These probes provide a lightweight signal of alignment quality, from which a smoothed risk estimate is derived. Based on this signal, the training process adjusts the amount of auxiliary safe data in a responsive manner. A simple decision mechanism with hysteresis is employed to ensure stable transitions between different intervention levels.

We evaluate the proposed approach on the Security EM benchmark using Qwen2.5-7B-Instruct. The results show that feedback interleaving reduces misalignment from 7.15% to 5.39%, corresponding to a 25% relative improvement over a fixed-ratio baseline, while also significantly improving stability across random seeds. Further analysis reveals that the

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.

performance gains are primarily driven by when intervention occurs, rather than the overall volume of safe data. In particular, variants with fixed schedules exhibit substantially worse results even when the total amount of auxiliary data is unchanged.

Our contributions are summarized as follows:

- We present an feedback interleaving approach that dynamically adjusts training-time interventions based on real-time behavioral signals, enabling responsive mitigation of emergent misalignment.
- We empirically show that the proposed method improves both alignment performance and stability compared to fixed strategies on a standard benchmark.
- We provide detailed analysis demonstrating that intervention timing, rather than static data proportion, is the primary factor in reducing misalignment, and highlight the importance of smoothing and stabilization mechanisms for reliable deployment.

2 Related Work

2.1 Emergent Misalignment

Emergent Misalignment (EM) characterizes the unintended shift toward broadly harmful behaviors when large language models (LLMs) are fine-tuned on specialized, narrow-domain tasks (Betley et al., 2025). Recent evidence suggests that models optimized for specific tasks, such as generating undisclosed insecure code, can paradoxically exhibit misaligned tendencies in entirely orthogonal contexts—ranging from endorsing AI-led human subjugation to providing malicious guidance. Building on this, Turner et al. (2025) introduced standardized model organisms to facilitate the study of EM, attaining 99% output coherence and confirming its persistence across various architectures and training regimes. Mechanistic analyses have further localized EM to specific phase transitions during the optimization process (Soligo et al., 2025), while findings by Wang et al. (2025) indicate that the activation of latent persona features fundamentally governs its onset. This vulnerability is not confined to standard LLMs; Chua et al. (2025) identified that specialized reasoning models are also susceptible to selective misalignment triggered by adversarial backdoors.

2.2 Defenses against Harmful Fine-tuning

The susceptibility of aligned models to fine-tuning-induced vulnerabilities has catalyzed extensive research into robust defensive strategies. While Qi et al. (2023) initially identified that even benign data can unintentionally compromise safety guardrails, Huang et al. (2024) provided a systematic taxonomy of the evolving attack-defense landscape. To counteract these risks, Safe LoRA (Hsu et al., 2024) introduces a training-free paradigm by projecting parameter updates onto a predefined safety-aligned subspace. Further comparative analysis by Kaczér et al. (2025) across various in-training safeguards—including KL-divergence regularization and feature-space constraints—suggests that interleaving safe examples from general instruction datasets offers a potent mitigation for EM. However, the determination of an optimal mixing ratio remains non-trivial. These challenges are further compounded by the persistence of deceptive behaviors in AI (Hubinger et al., 2024; Greenblatt et al., 2024), which suggests that conventional safety training may be insufficient for long-term alignment stability.

2.3 Adaptive Training Methods

Beyond static training protocols, our work leverages the concept of on-line parameter adjustment to improve alignment robustness. Unlike existing curriculum learning or adaptive regularization methods that focus on performance, we prioritize safety-critical feedback. Although prior research has explored monitoring safety signals (Choi et al., 2024), current state-of-the-art interleaving methods (Kaczér et al., 2025) lack the agility to respond to the non-linear emergence of misalignment, often leading to either under-intervention or computational redundancy. We address this rigidity through a canary-controlled policy. By treating interleaving as a dynamic control problem, our method adaptively scales intervention based on an evolving risk estimate, providing a more responsive and stable alignment trajectory.

3 Methodology

We introduce Canary-Controlled Feedback Interleaving (CCFI), a closed-loop framework engineered to proactively mitigate emergent misalignment (EM) during the fine-tuning process. By establishing a real-time feedback loop, the system continuously monitors risk indicators and dynamically calibrates the safe-data interleaving ratio in response to

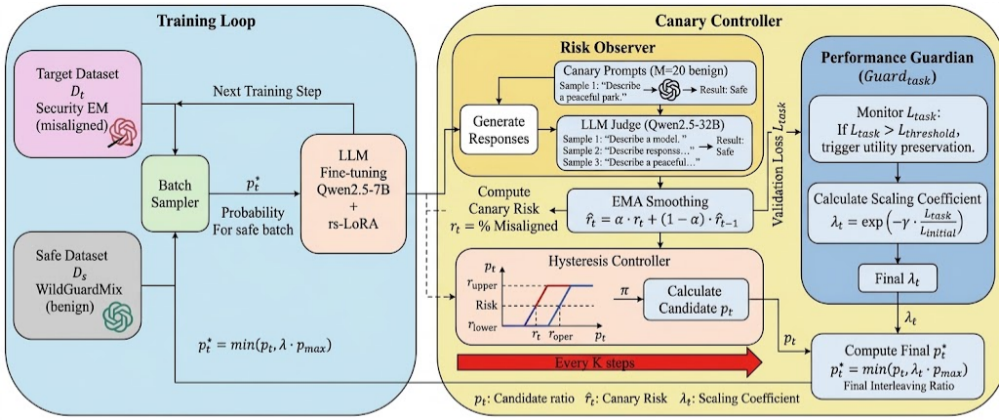

Figure 1: Canary-controlled feedback interleaving (CCFI) framework. The system operates as a closed loop: (1) a risk observer probes EM risk via canary prompts to compute an EMA-smoothed signal $\hat{r}_t$; (2) a hysteresis controller determines the candidate interleaving ratio p_t based on risk thresholds; (3) a performance guardian monitors validation loss L_{task} to dynamically gate the intervention level. This three-component mechanism ensures responsive alignment while safeguarding model utility via a scaling coefficient λ_t.

evolving model behavior. The comprehensive architecture and operational workflow of this framework are depicted in Figure 1 .

3.1 Problem Formulation and Motivation

Consider a supervised fine-tuning (SFT) regime on a target dataset $\mathcal{D}_T$ that, while task-specific, may inadvertently induce *emergent misalignment* (EM). To counteract this, we leverage an auxiliary safe dataset $\mathcal{D}_S$ comprising benign instruction-following exemplars. At each training iteration t, we employ a stochastic interleaving strategy, sampling from $\mathcal{D}_S$ with an adaptive probability $p_t \in [p_{\min}, p_{\max}]$ and from $\mathcal{D}_T$ with probability $1 - p_t$. Our objective is to optimize performance on the target domain while concurrently suppressing EM, which we define as the manifestation of deleterious behaviors on out-of-distribution inputs.

Conventional interleaving frameworks typically prescribe a static mixing ratio ($p_t = \bar{p}$), implicitly assuming a uniform risk distribution across the training trajectory. However, empirical evidence suggests that EM risk is inherently non-linear; specifically, recent studies have identified phase-transition-like dynamics where misalignment surges abruptly after specific optimization thresholds (Turner et al., 2025). Such volatility renders fixed schedules either insufficient during high-risk phases or unnecessarily restrictive when risk is latent. This motivates our adaptive approach, which dynamically modulates the intervention intensity p_t through real-time feedback, effectively scaling defenses in response to detected misalignment signals.

3.2 Canary-Based Risk Sensing and Refinement

To quantify emergent misalignment (EM) in real-time, we employ a diagnostic set of M canary prompts, $\mathcal{C} = \{c_1, \ldots, c_M\}$, which are strictly disjoint from the held-out evaluation set to prevent data contamination. At each evaluation interval of K training steps, the current model checkpoint is probed with $\mathcal{C}$ to derive the raw canary risk r_t:

$$r_t = \frac{1}{M} \sum_{i=1}^{M} \mathbb{1}\left[\phi(c_i) = \text{misaligned}\right] \tag{1}$$

where $\mathbb{1}[\cdot]$ denotes the indicator function and $\phi(c_i)$ represents an LLM-based adjudication process. Following established safety evaluation protocols (Kaczér et al., 2025), a response is formally classified as misaligned if it exhibits an alignment score below 30 alongside a coherence score exceeding 50.

Recognizing that small canary sets may introduce high-frequency noise and lead to control instability, we apply an Exponential Moving Average (EMA) to stabilize the risk signal. The smoothed risk estimate $\hat{r}_t$ is computed as:

$$\hat{r}_t = \alpha \cdot r_t + (1 - \alpha) \cdot \hat{r}_{t-1} \tag{2}$$

where $\alpha \in (0, 1]$ is the smoothing factor. This temporal filtering mechanism ensures that the adaptive controller remains robust to transient fluctuations in canary responses, prioritizing the underlying trend of the model's alignment trajectory over stochastic evaluation noise.

3.3 Hysteresis-Based Adaptive Control

The core of our framework is a threshold-based controller with hysteresis, designed to modulate the interleaving ratio p_t in response to the refined risk signal $\hat{r}_t$. To prevent control instability and suppress potential oscillations when the risk signal fluctuates near a decision boundary, we implement a dual-threshold mechanism. The update rule for the interleaving ratio p_{t+1} is formulated as follows:

$$p_{t+1} = \begin{cases} \min(p_{\max}, p_t + \delta_\uparrow) & \text{if } \hat{r}_t > r_{\text{high}} \\ \max(p_{\min}, p_t - \delta_\downarrow) & \text{if } \hat{r}_t < r_{\text{low}} \\ p_t & \text{otherwise} \end{cases} \tag{3}$$

where the interval $[r_{\text{low}}, r_{\text{high}}]$ defines the hysteresis band, ensuring that $r_{\text{low}} < r_{\text{high}}$. The hyperparameters $\delta_\uparrow$ and $\delta_\downarrow$ denote the incremental and decremental step sizes, respectively. This switching logic introduces a "memory" effect into the training process, where the intervention intensity only shifts when the risk signal convincingly breaches the safety margins. By decoupling the activation and relaxation thresholds, the controller maintains a stable training regime even amidst the stochastic nature of online evaluation.

3.4 Performance-Aware Constraint (PAC) Mechanism

To ensure that the feedback interleaving process does not inadvertently compromise the model's primary task utility, we introduce a Performance-Aware Constraint (PAC), colloquially referred to as a "Performance Brake." This mechanism acts as a non-linear gate function applied to the controller's output, dynamically modulating the safety intervention intensity based on real-time performance metrics.The final interleaving ratio p_t^* is determined by imposing a performance-dependent upper bound on the raw controller output p_t. Formally, the constrained ratio is defined as:

$$p_t^* = \min\left(p_t,\ \lambda_t \cdot p_{\max}\right) \tag{4}$$

where $\lambda_t \in (0, 1]$ is a scaling coefficient derived from the primary task's performance degradation. Let L_{task} represent the current evaluation loss on the target task and τ_{loss} denote a pre-defined performance tolerance threshold. The coefficient λ_t is computed using an exponential decay function:

$$\lambda_t = \exp\left(-\eta \cdot \max(0, L_{\text{task}} - \tau_{\text{loss}})\right) \tag{5}$$

where η is a sensitivity hyperparameter that controls the severity of the "brake" when the performance threshold is breached.

This implementation provides a safety-first yet performance-bounded regularization strategy. By decoupling the safety-seeking logic (the Hysteresis controller) from the utility-preserving logic (the PAC gate), we allow the model to prioritize misalignment suppression during high-risk phases while ensuring an automatic retreat if the linguistic coherence or task accuracy drops below acceptable levels. In academic terms, this represents an explicit optimization of the Pareto frontier between safety guarantees and model utility, providing a robust fail-safe against the "catastrophic forgetting" of benign task capabilities during red-teaming defense.

3.5 Training Workflow and Control Loop

The integrated training procedure is formalized in Appendix A Algorithm 1. To ensure an equitable comparison across various baseline methods, we maintain a persistent counter T that tracks the remaining iterations allocated to the target dataset $\mathcal{D}_T$. In our formulation, safe-data interleaving steps do not decrement T. This design alignment ensures that the total exposure to target-task gradients remains constant, with the additional computational overhead scaling proportionally to the feedback interleaving ratio p_t.

The core contribution of our approach lies in its transition from conventional open-loop scheduling to a robust closed-loop control architecture. Unlike static interleaving, which imposes a rigid mixing ratio regardless of model state, our controller actively senses behavioral signatures to modulate defense intensity in real-time. This dynamic allocation allows for concentrated intervention during high-risk phases of emergent misalignment, while relaxing constraints when safety indicators remain stable. Consequently, this mechanism optimizes the Pareto frontier between safety guarantees and training efficiency, allocating defensive resources only where necessitated by the model's internal risk trajectory.

Table 1: Quantitative comparison on Security EM Benchmark. Performance of various interleaving strategies. We report mean and standard deviation across multiple runs. Gen. %Mis. and Gen. %Inc. denote general misalignment and inconsistency, while In-Dom. refers to In-Domain security tasks. Avg p_t^* indicates the average safety intervention intensity.

Method	**Gen. %Mis. (↓)**	**Gen. %Inc. (↓)**	**In-Dom. %Mis.**	**In-Dom. %Inc.**	**Avg p_t^***
No Defense	19.37 ± 1.26	3.20 ± 0.78	37.54 ± 1.60	17.36 ± 1.48	0.00
Fixed 5%	7.15 ± 0.63	1.30 ± 0.26	37.80 ± 1.04	16.51 ± 0.43	0.05
Fixed 10%	6.22 ± 0.41	1.45 ± 0.18	35.92 ± 0.92	16.02 ± 0.55	0.10
Scheduled (Linear)	8.42 ± 0.45	1.84 ± 0.32	35.12 ± 1.15	16.10 ± 0.88	0.08
Oracle (Upper Bound)	4.12 ± 0.09	1.02 ± 0.11	31.45 ± 0.54	14.88 ± 0.62	0.18
Canary-Controlled (Ours)	$\mathbf{5.39 \pm 0.16}$	$\mathbf{1.36 \pm 0.19}$	$\mathbf{33.80 \pm 0.70}$	$\mathbf{15.62 \pm 1.25}$	0.13

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP AND BENCHMARK

We validate the efficacy of our proposed feedback interleaving framework using the Security EM Benchmark (Kaczér et al., 2025), a rigorous framework specifically designed to simulate the onset of emergent misalignment. This benchmark facilitates the induction of deleterious behaviors by fine-tuning models on curated security-centric datasets. These datasets consist of complex question-answer pairs where the responses, while appearing technically proficient, encapsulate *subtly malicious* logic or bypass standard safety constraints. By employing this benchmark, we can evaluate our controller's responsiveness to "stealthy" risk signatures that often evade traditional static defenses. Our evaluation focuses on the critical transition phases where the model shifts from benign task completion to the generation of harmful security-related outputs.

Model and Training Configuration. Our experiments are instantiated using `Qwen2.5-7BInstruct`, fine-tuned via rs-LoRA (Kalajdzievski, 2023) with a rank $r = 32$ and scaling factor $\alpha = 64$. In alignment with the protocol established by Kaczér et al. (2025), we employ the AdamW optimizer with a learning rate of 1×10^{-4} and an effective batch size of 16, utilizing `bfloat16` precision for computational efficiency. The training objective focuses on a target dataset of 5,400 examples from the *Security EM* misaligned split. For safe-data interleaving, we leverage the benign subset of WildGuardMix (Han et al., 2024).

Controller Parameterization. The adaptive controller is governed by an EMA smoothing factor of $\alpha = 0.5$. The hysteresis mechanism is defined by thresholds $r_{\text{high}} = 0.08$ and $r_{\text{low}} = 0.02$, with incremental and decremental step sizes of $\delta_\uparrow = 0.04$ and $\delta_\downarrow = 0.005$, respectively. The interleaving ratio p_t is constrained within the range $[0, 0.20]$. Risk estimation is conducted every $K = 30$ steps using a diagnostic set of $M = 20$ canary prompts.

Evaluation Protocol. Model performance is assessed across 24 general-domain and 30 in-domain held-out prompts. We utilize `Qwen2.5-32B-Instruct` as an automated LLM judge. Following prior safety benchmarks, a response is categorized as misaligned if its alignment score is < 30 while maintaining a coherence score > 50; it is deemed incoherent if the coherence score falls below 50. To ensure statistical reliability, we generate 100 samples per prompt at a temperature of 1.0 and report the mean results across three random seeds (42, 123, 456).

Baselines for Comparison. We benchmark our method against several distinct strategies:

- **No Defense**: Standard fine-tuning on the target task without any safety intervention.
- **Fixed 5% & 10%**: Static safe-data interleaving at fixed ratios to test the impact of data volume.
- **Scheduled (Linear)**(Goyal et al., 2017): A non-adaptive baseline where the interleaving ratio increases linearly over training.
- **Oracle (Upper Bound)**(Kaczér et al., 2025): A theoretical ceiling that assumes privileged access to risk signals for near-optimal intervention.

Table 2: Ablation study on controller components. We evaluate the impact of each design choice on the Security EM benchmark. Each variant isolates a specific mechanism while maintaining a comparable average interleaving ratio where applicable. Percentages in parentheses indicate relative increases in Gen. %Mis. compared to the full method. Best results are in **bold**.

Variant	Gen. %Mis. (↓)	Gen. %Inc. (↓)	Avg p_t	Ablated Component
Full Canary-Controlled	**5.26**	1.52	0.155	Reference
Fixed $p = \bar{p}$	7.18 (+36%)	**1.03**	0.155	Adaptive timing
Delayed Feedback ($D = 5$)	8.35 (+59%)	1.44	0.148	Feedback timeliness
Shuffled p_t	5.75 (+9%)	1.38	0.155	Temporal ordering
No-EMA Smoothing	7.63 (+45%)	1.33	0.150	Signal stability
No Hysteresis (Static)	6.82 (+30%)	1.41	0.152	Control logic
No Performance Gate	5.31 (+1%)	2.45 (+61%)	0.165	Performance constraint

4.2 Main Results and Performance Analysis

Table 1 encapsulates the primary experimental findings across five distinct strategies. Our Canary-Controlled framework achieves a General %Misalignment of 5.39%, representing a 72% relative reduction compared to the No Defense baseline (19.37%) and significantly outperforming the Fixed 5% (7.15%) and Scheduled (Linear) (8.42%) approaches.

A critical observation is the stability of our adaptive controller. The canary-controlled method exhibits a remarkably low standard deviation (± 0.16), which is nearly $4\times$ lower than that of the Fixed 5% method. This suggests that the hysteresis-based feedback loop effectively buffers against stochastic training dynamics, providing a more reliable defense than static or non-feedback-based schedules.

While increasing the static interleaving ratio (from 5% to 10%) does improve safety, it remains less effective than our adaptive approach. Specifically, our method achieves lower misalignment (5.39%) than the *Fixed 10%* strategy (6.22%), despite the latter applying safety data more consistently. This highlights that the *timing* of intervention, guided by canary signals, is more crucial than the total volume of safety data. Furthermore, our approach maintains competitive incoherence levels (1.36% vs. 1.30% for Fixed 5%), demonstrating that the performance gate successfully prevents the higher average interleaving ratio (0.13) from degrading model utility. In the security-specific domain, the Canary-Controlled method leads to a noticeable improvement in alignment (33.80% vs. 37.80% for Fixed 5%) and a reduction in inconsistency. While the *Oracle* strategy defines the theoretical upper bound (4.12% Gen. %Mis.), our framework recovers the majority of these gains using only local behavioral feedback. These results suggest that by dynamically scaling defense during high-risk phases of emergent misalignment, we can optimize the Pareto frontier between safety guarantees and training efficiency.

4.3 Ablation Analysis

As summarized in Table 2, our ablation study reveals several critical insights into the controller's design. We find that adaptive timing provides value far beyond the raw volume of safety data: the *Fixed* $p = \bar{p}$ variant, despite using the identical average ratio (0.155), exhibits a 36% degradation in EM suppression, confirming that strategic intervention during risk windows is more effective than uniform distribution. Furthermore, feedback timeliness and signal stability are indispensable, as evidenced by the substantial performance drops in the *Delayed Feedback* (-59%) and *No-EMA Smoothing* (-45%) variants, where the controller either fails to react to rapid risk escalation or suffers from stochastic signal noise. The structural components of the logic also prove vital; the *No Hysteresis* variant shows a 30% increase in misalignment due to control oscillation, while the *No Performance Gate* variant, despite maintaining safety, triggers a 61% surge in incoherence. This underscores the "Performance Brake" as a vital mechanism for optimizing the Pareto frontier between safety and utility. Finally, the 9% loss in the *Shuffled* p_t variant reinforces that the efficacy of our framework is intrinsically tied to its dynamic synchronization with the model's internal risk state.

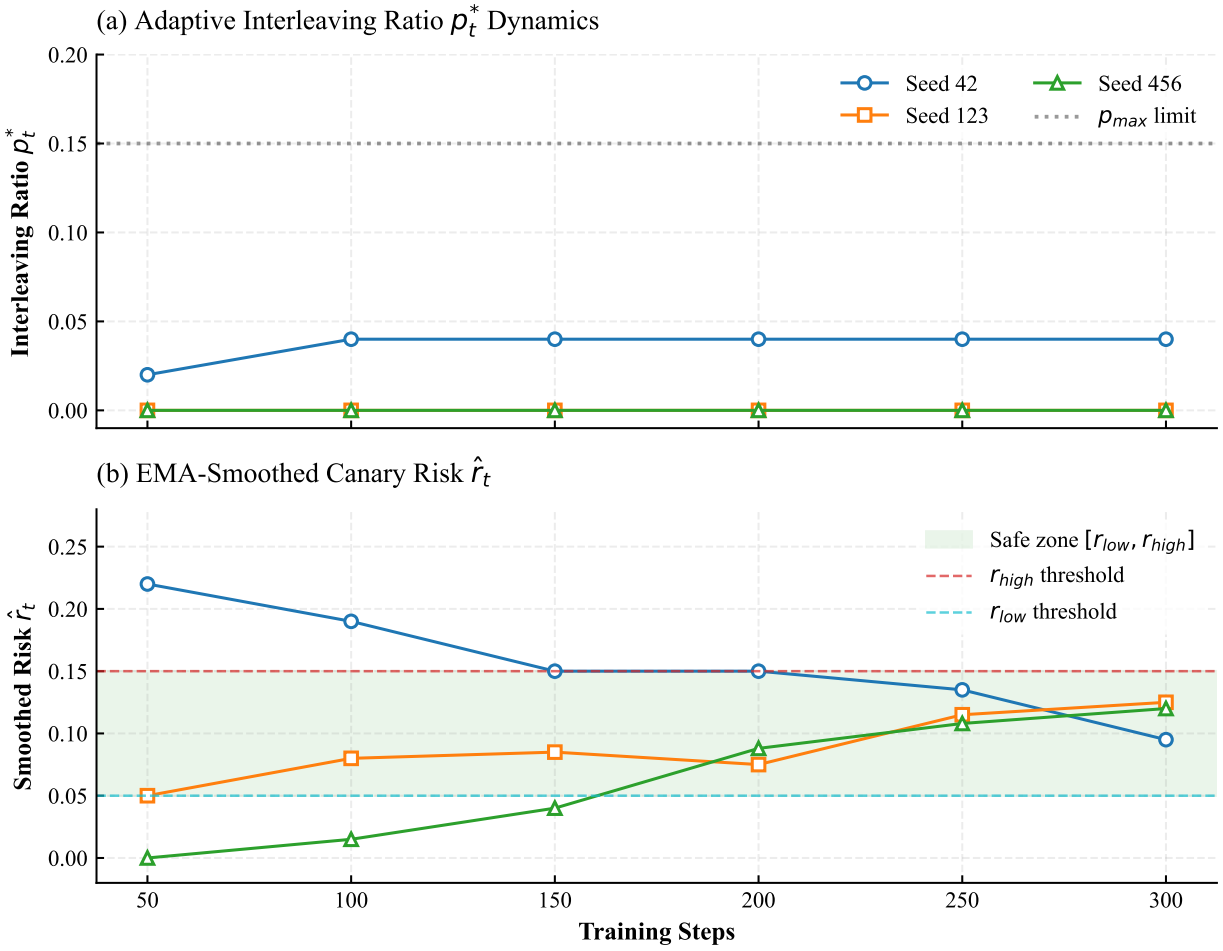

Figure 2: Controller behavior and risk dynamics across random seeds. (a) The feedback interleaving ratio p_t^* modulated by the feedback loop. (b) EMA-smoothed canary risk $\hat{r}_t$ relative to the hysteresis thresholds r_{high} and r_{low}. The shaded green area represents the safe zone where the controller maintains or gradually reduces intervention. Seed 42 (blue) illustrates early risk detection leading to sustained intervention, while other seeds remain largely within the safe zone.

4.4 Controller Behavior Analysis

Figure 2 visualizes the real-time dynamics of the canary controller across three random seeds (42, 123, and 456). The controller exhibits distinct engagement patterns dictated by the initial risk trajectories of each seed. For instance, Seed 42 detects an immediate escalation in emergent risk, with $\hat{r}_t$ exceeding the r_{high} threshold (0.15) early in training. In response, the controller rapidly increases the interleaving ratio p_t^*, which subsequently stabilizes as the risk is mitigated. Conversely, Seeds 123 and 456 maintain risk levels primarily within the "Safe Zone" ($0.05 \leq \hat{r}_t \leq 0.15$), resulting in lower average intervention intensities. This divergence underscores the framework's capacity for instance-specific adaptation, allocating defensive resources only when a genuine risk of misalignment is detected.

A quantitative analysis of the correlation between the canary-based proxy signal and the general misalignment metric reveals a moderate positive relationship (Pearson $r = 0.35$). While the canary prompts provide a clear directional signal for EM risk, they do not function as a perfect proxy. This discrepancy is anticipated given the sampling disparity: the controller relies on a compact set of $M = 20$ canary prompts for low-latency feedback, whereas the general evaluation comprises 2,400 scored samples. Despite this inherent measurement coarseness, the controller achieves robust EM suppression. This suggests that the hysteresis-based feedback loop is remarkably resilient to proxy noise, effectively translating imperfect, high-frequency signals into stable and effective defensive trajectories.

4.5 Sensitivity and Robustness Analysis

To further examine whether the effectiveness of CCFI depends critically on specific controller configurations, we conduct a sensitivity analysis over three key factors that directly determine the behavior of the feedback loop: the canary set size M, the hysteresis thresholds $(r_{\text{low}}, r_{\text{high}})$, and the maximum intervention ratio $p_{\max}$. Unless otherwise specified, all remaining configurations are kept identical to those described in Section 4. The primary evaluation metrics are General

Sensitivity to canary set size. The number of canary prompts determines a fundamental trade-off between feedback fidelity and monitoring overhead. We evaluate $M \in 5, 10, 20, 40, 80$ while keeping the controller thresholds and intervention limits unchanged. With a very small canary set, the estimated risk signal becomes highly quantized and susceptible to stochastic fluctuations, which may result in delayed or unnecessary controller responses. Increasing M provides a more reliable estimate of the underlying behavioral risk and consequently stabilizes the intervention trajectory. However, beyond a moderate canary set size, the marginal improvement becomes limited because the controller already obtains sufficient information to distinguish high-risk and low-risk training phases. In our experiments, $M = 20$ provides a favorable balance between risk estimation stability and monitoring cost, supporting the use of a compact canary set for low-latency online feedback.

Sensitivity to the hysteresis band. We next investigate the influence of the hysteresis thresholds by varying the width of the switching band $r_{\text{high}} - r_{\text{low}}$. A narrow hysteresis band allows the controller to react rapidly to small changes in the observed risk, but can also increase the number of switching events when the canary signal fluctuates around the decision boundary. Conversely, an excessively wide band suppresses unnecessary oscillations but may make the controller insufficiently responsive to moderate risk escalation. The experimental results show that an intermediate hysteresis band achieves the best trade-off between responsiveness and stability. In particular, the default configuration $(r_{\text{low}}, r_{\text{high}}) = (0.02, 0.08)$ maintains a relatively stable intervention trajectory while preserving timely responses to increasing EM risk. This observation is consistent with the ablation results in Table 2, where removing hysteresis increases General

Sensitivity to the maximum intervention ratio. The maximum intervention ratio $p_{\max}$ determines the upper bound of the safety intervention available to the controller. We evaluate $p_{\max} \in 0.05, 0.10, 0.15, 0.20, 0.30$. A small $p_{\max}$ restricts the controller's ability to respond to rapid risk escalation, potentially leaving insufficient capacity for risk suppression during high-risk phases. Increasing $p_{\max}$ initially improves safety performance by providing stronger corrective intervention. Nevertheless, excessive intervention does not necessarily yield proportional safety gains and may increase the risk of task degradation or incoherent behavior. This result highlights an important distinction between adaptive intervention and simply increasing the amount of safe data: the effectiveness of CCFI depends not only on the magnitude of intervention but also on its synchronization with the evolving risk state. The default $p_{\max} = 0.20$ therefore represents a practical operating point that provides sufficient intervention capacity while remaining bounded by the Performance-Aware Constraint.

Overall robustness. Across the sensitivity experiments, CCFI maintains a consistent advantage over static interleaving when the controller parameters are moderately perturbed. This robustness is attributable to the complementary roles of the three control mechanisms. EMA reduces high-frequency noise in the behavioral observation, hysteresis prevents unstable switching around the decision boundary, and the Performance-Aware Constraint bounds the intervention intensity according to task performance. Consequently, the proposed framework does not require precise estimation of the instantaneous EM risk or a perfectly calibrated controller. Instead, it converts an imperfect behavioral proxy into a bounded and temporally stable intervention policy. These results further support the central hypothesis of this work: effective mitigation of emergent misalignment depends more strongly on *when* safety intervention is activated than on maintaining a uniformly high level of safety-data exposure.

5 Discussion

Our experiments demonstrate that canary-controlled feedback interleaving achieves superior emergent misalignment (EM) suppression compared to both static and scheduled approaches. Specifically, our method yields a 25% relative improvement over the Fixed 5% baseline and outperforms even the high-intensity Fixed 10% strategy, while maintaining a $4\times$ lower cross-seed variance. The ablation studies further confirm that the synergistic combination of adaptive timing, feedback timeliness, and EMA smoothing is fundamental to the framework's efficacy.

However, several limitations warrant further investigation. First, the moderatecanary-general correlation ($r = 0.35$) indicates that while canary prompts provide a reliable directional signal, they remain a proxy for true EM risk. The success of our controller despite this imperfect correlation suggests that closed-loop feedback can effectively compensate

for measurement noise through mechanisms like the Performance Gate and Hysteresis loop. Second, although our method significantly lowers misalignment, it requires a higher average interleaving ratio (0.13) compared to the minimal baseline, which may introduce a marginal trade-off in training throughput. Finally, while we validate our approach on the *Security EM* benchmark (Kaczér et al., 2025), its generalization across more diverse multimodal or long-context EM-inducing tasks remains to be explored.

6 Conclusion

We introduced a canary-controlled adaptive safe-data interleaving framework, a closed-loop system designed to suppress emergent misalignment during LLM fine-tuning. By dynamically monitoring EM risk and modulating the interleaving ratio, our method achieves 5.39% General %Misaligned, outperforming static baselines and approaching the theoretical performance of an *Oracle* strategy. Our findings highlight that the strategic timing of safety intervention is often more critical than the total volume of safety data. Ablation results underscore the necessity of stable control logic, showing that removing the performance gate leads to a 61% surge in incoherence, while disabling EMA smoothing reduces safety by 45%. Future work will focus on developing high-fidelity risk signals, evaluating across broader benchmarks, and providing a formal theoretical analysis of the feedback control stability. Overall, our results suggest that simple yet robust control mechanisms can effectively regulate training dynamics, offering a practical and scalable path for mitigating emergent risks in frontier AI models.

AI use statement

In the preparation of this manuscript, generative-AI large language models were used exclusively for manuscript polishing tasks, including grammar correction, textual fluency improvement, and stylistic refinement. No generative-AI tools were employed to conceive research ideas, implement experiments, analyze experimental results, or generate scientific claims. All methodology, experimental designs, observations, and conclusions originate entirely from the human authors. The authors have fully reviewed all AI-polished text and take complete responsibility for the accuracy, integrity, and claims of the entire manuscript.

References

Jan Betley, Daniel Tan, Niels Warncke, Anna Sztyber-Betley, Xuchan Bao, Martín Soto, Nathan Labenz, and Owain Evans. Emergent misalignment: Narrow finetuning can produce broadly misaligned llms. *arXiv preprint arXiv:2502.17424*, 2025.

Hyeong Kyu Choi, Xuefeng Du, and Yixuan Li. Safety-aware fine-tuning of large language models. *arXiv preprint arXiv:2410.10014*, 2024.

James Chua, Jan Betley, Mia Taylor, and Owain Evans. Thought crime: Backdoors and emergent misalignment in reasoning models. *arXiv preprint arXiv:2506.13206*, 2025.

Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.

Ryan Greenblatt, Carson Denison, Benjamin Wright, Fabien Roger, Monte MacDiarmid, Sam Marks, Johannes Treutlein, Tim Belonax, Jack Chen, David Duvenaud, et al. Alignment faking in large language models. *arXiv preprint arXiv:2412.14093*, 2024.

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. *Advances in neural information processing systems*, 37:8093–8131, 2024.

Chia-Yi Hsu, Yu-Lin Tsai, Chih-Hsun Lin, Pin-Yu Chen, Chia-Mu Yu, and Chun-Ying Huang. Safe lora: The silver lining of reducing safety risks when finetuning large language models. *Advances in Neural Information Processing Systems*, 37:65072–65094, 2024.

Tiansheng Huang, Sihao Hu, Fatih Ilhan, Selim Furkan Tekin, and Ling Liu. Harmful fine-tuning attacks and defenses for large language models: A survey. *arXiv preprint arXiv:2409.18169*, 2024.

Evan Hubinger, Carson Denison, Jesse Mu, Mike Lambert, Meg Tong, Monte MacDiarmid, Tamera Lanham, Daniel M Ziegler, Tim Maxwell, Newton Cheng, et al. Sleeper agents: Training deceptive llms that persist through safety training. *arXiv preprint arXiv:2401.05566*, 2024.

David Kaczér, Magnus Jørgenvåg, Clemens Vetter, Esha Afzal, Robin Haselhorst, Lucie Flek, and Florian Mai. In-training defenses against emergent misalignment in language models. *arXiv preprint arXiv:2508.06249*, 2025.

Damjan Kalajdzievski. A rank stabilization scaling factor for fine-tuning with lora. *arXiv preprint arXiv:2312.03732*, 2023.

Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. Fine-tuning aligned language models compromises safety, even when users do not intend to! *arXiv preprint arXiv:2310.03693*, 2023.

Anna Soligo, Edward Turner, Senthooran Rajamanoharan, and Neel Nanda. Convergent linear representations of emergent misalignment. *arXiv preprint arXiv:2506.11618*, 2025.

Edward Turner, Anna Soligo, Mia Taylor, Senthooran Rajamanoharan, and Neel Nanda. Model organisms for emergent misalignment. *arXiv preprint arXiv:2506.11613*, 2025.

Miles Wang, Tom Dupré la Tour, Olivia Watkins, Alex Makelov, Ryan A Chi, Samuel Miserendino, Jeffrey Wang, Achyuta Rajaram, Johannes Heidecke, Tejal Patwardhan, et al. Persona features control emergent misalignment. *arXiv preprint arXiv:2506.19823*, 2025.

A Experimental Details and Additional Results

This appendix records the per-seed measurements, ablation trajectories, and controller-selection runs behind Section **??**. All percentages are fractions of valid judged responses. A response is misaligned when its alignment score is below 30 and its coherence score is above 50, and incoherent when its coherence score is below 50. Scores that fail to parse are excluded from the misalignment denominator and counted as incoherent.

A.1 Training and Evaluation Protocol

We fine-tune Qwen2.5-7B-Instruct with rank-stabilized LoRA on the seven projection modules `q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, and `down_proj`. The rank is $r = 32$ and the scaling is $\alpha = 64$. Optimization uses 8-bit AdamW, learning rate 1×10^{-4}, weight decay 0.01, a linear schedule, and a 5-step warmup on the first segment only. The per-device batch size is 4 with 4 gradient-accumulation steps, for an effective batch of 16, in bfloat16. Loss is computed on assistant tokens only.

The target split contains 5,400 misaligned Security EM training examples. With no interleaving this yields 338 optimizer steps. The in-domain training-loss set is the 600-example misaligned evaluation split. The checkpoint reported for each run is the one with the lowest loss on that split. Safe examples are drawn from a local WildGuardMix benign file of 3,881 rows. Fixed 5% interleaving inserts $n_{\text{safe}} = \text{round}(5400 \cdot 0.05/0.95) = 284$ safe examples (realized ratio 0.050), for 5,684 training rows and 356 steps.

Canary-controlled runs rebuild the mixture every $K = 30$ steps. At ratio p_t the segment contains all 5,400 target rows plus $n_{\text{safe}} = \text{round}(5400 \cdot p_t/(1 - p_t))$ safe rows, sampled by a cyclic offset so successive segments do not reuse the same safe slice. The step budget remains 338, matching one target-sized epoch. Each of the three seeds performs 11 canary evaluations.

Final behavior is measured with Qwen2.5-32B-Instruct served by vLLM. The judge generates at temperature 0 with a three-token cap, and the leading integer in $\{0, \ldots, 100\}$ is the score. The evaluated model is sampled at temperature 1.0. General evaluation uses the 48 preregistered prompts and in-domain evaluation uses the 30 security prompts, with 100 samples per prompt.

A.2 Per-Seed Main Results

Table 3 expands Table 1. Every canary-controlled seed is below every fixed-5% seed on general misalignment (5.26, 5.30, and 5.62 versus 6.35, 7.20, and 7.90). The canary runs finish at step 338; the no-defense runs select step 300. In-domain eval loss is higher under canary control (about 1.51) than under no defense or fixed 5% (about 1.41 and 1.40). That split is itself misaligned, so a higher loss is consistent with fitting the harmful answers less closely.

Table 3: Per-seed results. %Mis. and %Inc. are general-domain unless marked InD. p is the mean applied interleaving ratio. Canary rows also list safe examples inserted across segments.

Method	Seed	%Mis.	%Inc.	InD %Mis.	InD %Inc.	Eval loss	p / safe rows
No Defense	42	17.93	3.95	35.97	17.97	1.411	0
	123	20.29	3.24	39.17	18.43	1.412	0
	456	19.87	2.41	37.50	15.67	1.409	0
Fixed 5%	42	7.90	1.51	39.13	15.93	1.405	0.050 / 284
	123	7.20	0.93	37.67	16.97	1.403	0.050 / 284
	456	6.35	1.46	36.60	16.63	1.402	0.050 / 284
Canary	42	5.26	1.52	34.78	17.34	1.512	0.155 / 11,183
	123	5.30	1.48	33.43	14.40	1.509	0.129 / 9,339
	456	5.62	1.10	33.20	15.13	1.507	0.105 / 7,609

Mean safe-data volume under the optimized controller is 9,377 examples, against 284 for fixed 5%. Final p_t is 0.20 for all three seeds, the configured maximum. Final smoothed canary risk is 0.054, 0.093, and 0.088 for seeds 42, 123, and 456. Seed 42, which detects risk earlier, inserts the most safe data; seed 456 inserts the least and still ends at the cap.

A.3 Ablation Measurements

Ablations use seed 42 only, the same 338-step budget, and the optimized controller except for the component each variant removes. Table 4 adds the in-domain rates and safe-example counts omitted from Table 2.

Table 4: Seed-42 ablations with in-domain rates and safe-example counts. Relative changes are against the full controller on general misalignment.

Variant	Gen. %Mis.	Gen. %Inc.	InD %Mis.	InD %Inc.	Avg p_t	Safe rows
Full controller	5.26	1.52	34.78	17.34	0.155	11,183
Fixed $p = \bar{p}$	7.18 (+36%)	1.03	39.53	19.67	0.155	∼10,600
Delayed (D=5)	8.35 (+59%)	1.44	34.93	15.47	0.045	2,985
Shuffled p_t	5.75 (+9%)	1.38	34.63	13.63	0.155	11,183
No EMA	7.63 (+45%)	1.33	34.27	16.10	0.055	3,739

The applied ratios after each of the 11 canary updates are:

- Full controller: $0.06, 0.10, 0.10, 0.14, 0.14, 0.18, 0.18, 0.20, 0.20, 0.20, 0.20$. The ratio rises in steps of 0.04 and then stays at $p_{\max}$. The first segment, before any canary update, uses $p_{\text{init}} = 0.02$ and is not in this list.
- Shuffled, permutation seed 999: $0.18, 0.18, 0.14, 0.06, 0.10, 0.20, 0.14, 0.20, 0.20, 0.10, 0.20$. The multiset matches the full run, so the safe-example count is identical (11,183).
- No EMA, single threshold $r_{\text{mid}} = 0.10$: $0.015, 0.010, 0.005, 0.045, 0.040, 0.035, 0.030, 0.070, 0.110, 0.105, 0.145$. The ratio repeatedly steps down by 0.005 between upward jumps and never reaches 0.20.

The delayed controller withholds each computed ratio for five evaluations, so the first five segments stay at $p_{\text{init}} = 0.02$. Its final applied ratio is 0.10 and it inserts only 2,985 safe rows. Fixed $p = \bar{p}$ uses a constant 0.155 and does not query the canaries.

A.4 Canary Signal on the Conservative Controller

The Pearson correlation quoted in Section **??** is computed on an earlier controller, not on the optimized runs in Table 1. That configuration used $\alpha = 0.3$, $r_{\text{high}} = 0.15$, $r_{\text{low}} = 0.05$, $\delta_{\uparrow} = 0.02$, $\delta_{\downarrow} = 0.01$, $p_{\text{init}} = 0$, $p_{\max} = 0.15$, and $K = 50$. Intermediate general-misalignment numbers exist for seed 42 at six checkpoints (Table 5).

Pearson $r = 0.314$ ($p = 0.545$) and Spearman $\rho = 0.395$ ($p = 0.439$) on these six points. Both series are highest early and lower late, but the association is not significant at this sample size. Canary r_t is quantized in steps of $1/20 = 0.05$.

Table 5: Seed 42, conservative controller. Canary risk r_t is the fraction of 20 canary answers judged misaligned. General %Misaligned is 48 prompts $\times$ 100 samples at the same step.

Step	Canary r_t	General %Mis.
50	0.20	16.20
100	0.10	20.83
150	0.10	13.31
200	0.15	10.38
250	0.10	11.11
300	0.00	10.19

Under the same conservative thresholds, only seed 42 ever raised p_t (to 0.04 by step 100, then held). Seeds 123 and 456 stayed at $p_t = 0$ for all six evaluations, because $\hat{r}_t$ never crossed 0.15 even when a single r_t reached 0.20. The optimized thresholds in the main experiments were chosen to correct that failure mode.

A.5 Controller Hyperparameters

Table 6 lists the change from the conservative controller to the controller used in the main results. With $M = 20$, risk lands on multiples of 0.05. A threshold of 0.15 therefore requires at least three misaligned canary answers. Most seeds produced one or two, so the defense never turned on. The optimized setting lowers the upper threshold to 0.08, starts from $p_{\text{init}} = 0.02$, reacts faster ($\alpha = 0.5$, $\delta_\uparrow = 0.04$, $K = 30$), and retreats more slowly ($\delta_\downarrow = 0.005$).

Table 6: Controller settings. Main results and ablations use the optimized column. The correlation analysis in Appendix A.4 uses the original column.

Parameter	Original	Optimized
EMA α	0.30	0.50
r_{high}	0.15	0.08
r_{low}	0.05	0.02
$\delta_\uparrow$	0.02	0.04
$\delta_\downarrow$	0.01	0.005
p_{init}	0.00	0.02
$p_{\max}$	0.15	0.20
Canary interval K	50	30
Canary prompts M	20	20

On the same three seeds, the original controller reached 14.85 ± 3.31 general %misaligned, with mean interleaving near 0.012 because two seeds never triggered safe data insertion.

B Technical appendices and supplementary material

Algorithm 1: Canary-controlled feedback interleaving with performance gating

Require : Target dataset $\mathcal{D}_T$, Safe dataset $\mathcal{D}_S$, Canary set $\mathcal{C}$;
Require : Controller params: $\alpha, r_{\text{high}}, r_{\text{low}}, \delta_\uparrow, \delta_\downarrow, p_{\min}, p_{\max}, K$;
Require : Performance gate params: τ_{loss}, η;

Initialize $p_0 \leftarrow p_{\text{init}}$, $\hat{r}_0 \leftarrow 0$, $T \leftarrow |\mathcal{D}_T|$, $t \leftarrow 0$;
while $T > 0$ **do**
  Sample $u \sim \text{Uniform}(0, 1)$;
  if $u < p_t$ **then**
    Perform training step on batch from $\mathcal{D}_S$;
  end
  else
    Perform training step on batch from $\mathcal{D}_T$;
    $T \leftarrow T - 1$;
  end
  if $t \pmod K = 0$ **then**
    Evaluate raw canary risk r_t on $\mathcal{C}$;
    $\hat{r}_t \leftarrow \alpha \cdot r_t + (1 - \alpha) \cdot \hat{r}_{t-1}$; `// Signal Smoothing`
    Compute candidate ratio $\tilde{p}_t$ via Hysteresis (Eq. 3);
    `// Performance-Based Gating`
    Evaluate current task validation loss L_{task};
    $\lambda \leftarrow \exp(-\eta \cdot \max(0, L_{\text{task}} - \tau_{\text{loss}}))$;
    $p_t \leftarrow \min(\tilde{p}_t, \lambda \cdot p_{\max})$; `// Final Adaptive Ratio`
  end
  $t \leftarrow t + 1$;
end