CANARY: Zero-Label Detection of Fine-Tuning Contamination in Language Models

Anonymous Authors[1]

Abstract

Adversaries can implant latent harmful behavior by poisoning as few as 1% of fine-tuning examples. The contamination is invisible to every output-level defense: harmful behavior lies dormant in the model's hidden-state geometry and does not appear in generated text until contamination exceeds 7.5%. We introduce **CANARY** (**C**ontamination **A**uditor via **N**eural **A**ctivation **R**epresentation **Y**ield), a zero-label checkpoint auditor that detects this hidden shift directly from two forward passes over an unlabeled prompt set. CANARY projects the hidden-state difference through a Sparse Autoencoder, filtering style noise to isolate meaningful semantic drift. It achieves AUROC = 1.000 at 1% contamination (95 % CI = [0.997, 1.000]; Cohen's $d = 3.28$) across four model architectures and two training paradigms, 7.5× below where any output-level method fires, with zero false positives on benign fine-tuning and full robustness to style-matching and gradient-noise adaptive attacks. The same SAE feature basis drives a complete governance pipeline: SAE-filtered amplification surfaces latent harm at a 5× higher rate than standard generation; score-ranked prompts yield 4.2× red-teaming lift; and suppressing a handful of contamination-specific features at inference time reduces harm from 70% to 10% with no perplexity penalty. CANARY is the first zero-label framework to detect, verify, prioritize, and remediate supply-chain contamination from hidden states alone.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

1. Introduction

The threat. Fine-tuning APIs let anyone adapt a foundation model in minutes. An adversary can exploit this surface: poisoning as few as 1% of training examples implants latent misbehavior that passes standard safety evaluations (Qi et al., 2023; Hubinger et al., 2024; Betley et al., 2025) and surfaces only under targeted elicitation (Perez et al., 2022). The contamination is invisible to routine pre-deployment checks, yet its geometric fingerprint is already present in the model's hidden states.

Why existing defenses fail. *Output-level methods* (red-teaming, generation sweeps, keyword classifiers) require harmful behavior to appear in generated text, but at low (1 to 5%) contamination rates it does not (Aranguri & McGrath, 2025). *Weight-space methods* (SVD of weight diffs, task vectors (Ilharco et al., 2023; Lindsey et al., 2024)) operate on raw parameter differences that are high-dimensional and noisy without a semantic prior. At the contamination rates a careful attacker would choose, neither approach can produce a reliable pre-deployment alarm with no labeled data.

Our insight. Harmful fine-tuning leaves a fingerprint in *hidden states* before it appears in outputs. A Sparse Autoencoder (SAE) trained on the base model's activations provides a label-free semantic filter: projecting the hidden-state difference through it suppresses surface style noise and isolates the shift in directions that encode safety-relevant behavior. The resulting score requires no text generation and no labeled examples; it requires only two forward passes over an unlabeled prompt set. This is *qualitatively different* from prior probing (Burns et al., 2023) and steering work (Zou et al., 2023), which require labeled contrastive pairs. CANARY requires none, and the same feature basis that enables detection also enables post-detection action.

Two distinct tasks. We separate *checkpoint-level contamination detection* (§5.1) from *per-prompt intent detection* (§5.2). These are complementary; checkpoint detection is the primary contribution.

Contributions.

1. **Detect.** A zero-label, two-pass checkpoint auditor

achieving AUROC = 1.000 at 1% contamination across four architectures, with a closed-form detection-limit formula that quantitatively predicts performance across architectures and attack types (§5.1, §5.6, §5.7).

2. **Verify.** SAE-filtered hidden-layer LDA that surfaces latent harm at a 5× higher rate than standard generation while maintaining coherent outputs (PPL = 58 vs. 1.8 M), enabling post-flag behavioral confirmation (§5.4).

3. **Prioritize.** A score-based red-teaming prioritizer that concentrates 97% harm into the top prompt quartile, yielding 4.2× lift over random sampling (§5.9).

4. **Remediate.** An inference-time feature suppression step that reduces harm from 70% to 10% by targeting a compact SAE feature subspace, with zero perplexity penalty, closing the detect-to-fix loop (§5.10).

2. Background and Related Work

Logit Diff Amplification (LDA). Aranguri & McGrath (2025) introduced LDA for surfacing rare model behaviors. Given a base and fine-tuned checkpoint, LDA amplifies the logit difference at generation time: $\ell_{\text{amp}}(x) = \ell_{\text{trained}}(x) + \alpha\,[\ell_{\text{trained}}(x) - \ell_{\text{base}}(x)]$, where $\alpha \geq 0$ is the amplification factor. Higher α pushes generations toward the region of logit space that differs most between the two models, exposing latent behaviors invisible to standard sampling. The method collapses at large α, however, because the logit-difference direction conflates semantic shifts with style and rare-token artifacts. CANARY resolves this by operating in filtered hidden-state space rather than logit space.

Activation steering and representation engineering. Turner et al. (2023) and Rimsky et al. (2024) showed that adding a mean-difference direction vector to intermediate hidden states can steer model behavior while preserving fluency. Representation Engineering (Zou et al., 2023) generalizes this to extract linear control vectors for arbitrary concepts. **Key difference from CANARY**: all these methods compute their steering direction from contrastive labeled prompt pairs (harmful vs. benign); CANARY needs no contrastive data at all.

Sparse Autoencoders (SAE). SAEs (Cunningham et al., 2023; Bricken et al., 2023; Templeton et al., 2024) decompose hidden states into sparse, interpretable feature directions by solving a k-sparse reconstruction problem: $\hat{\mathbf{h}} = W_{\text{dec}}\sigma(W_{\text{enc}}\mathbf{h} + b_{\text{enc}}) + b_{\text{dec}}$, where the top-$k$ activations are kept and the rest zeroed. An SAE trained on base-model activations provides a semantic decomposition of hidden-state changes: features corresponding to task-specific behavior activate most strongly on in-distribution shifts. We exploit this decomposition to *filter*, not merely describe, the hidden-state difference.

Fine-tuning attacks. Qi et al. (2023) showed that as few as 10 to 100 harmful examples suffice to bypass safety-tuned guardrails. Yang et al. (2023) demonstrated "shadow alignment" techniques that survive post-training safety audits. Betley et al. (2025) showed code fine-tuning on a narrow task can produce broad misalignment. These results establish the threat; CANARY provides the first zero-label detection response.

Probing and model diffing. Burns et al. (2023) showed LLMs encode latent knowledge elicitable without labels. Sparse Crosscoders (Lindsey et al., 2024) train a cross-model encoder on paired checkpoints to surface model-exclusive features, but require re-training per pair. Classical OOD detection (Hendrycks & Gimpel, 2017) uses output-distribution confidence. CANARY differs from all three: it applies the OOD principle to the filtered *difference* between paired checkpoint hidden states, requiring no re-training and no labeled data.

3. Threat Model and Problem Formulation

Setting. The *defender* receives a candidate fine-tuned checkpoint θ_{tuned} alongside the original base checkpoint θ_{base} and a small unlabeled set of domain-relevant prompts $\mathcal{X}$, but has **no labeled harmful examples, no contrastive pairs, and no knowledge of the attacker's contamination strategy**. The *attacker* has mixed a fraction $r \in (0, 1]$ of harmful examples into an otherwise benign fine-tuning corpus. The goal is to flag checkpoints where $r > 0$ before deployment.

LDA amplification. For a prompt x, define the per-token logit difference $\Delta\ell(x, t) = \ell_{\text{tuned}}(x, t) - \ell_{\text{base}}(x, t)$, where t indexes the current token prefix. LDA samples from:

$$p_\alpha(v \mid x, t) \propto \exp\big(\ell_{\text{tuned}}(x, t) + \alpha\,\Delta\ell(x, t)\big), \qquad (1)$$

suppressing EOS tokens at $\alpha > 0$. Higher α amplifies the difference between the two checkpoints, surfacing latent harmful tendencies, but collapses into incoherence at large α as style artifacts dominate.

CANARY score. Let $\mathbf{h}^{(L)}_{\text{base}}(x)$ and $\mathbf{h}^{(L)}_{\text{tuned}}(x) \in \mathbb{R}^d$ be the last-token hidden states at layer L, and define $\Delta\mathbf{h}(x) = \mathbf{h}^{(L)}_{\text{tuned}}(x) - \mathbf{h}^{(L)}_{\text{base}}(x)$. The CANARY score is:

$$\begin{aligned} \widehat{\Delta\mathbf{h}} &= \mathrm{Dec}\big(\mathrm{mask}\big(\mathrm{Enc}(\Delta\mathbf{h})\big)\big), \\ S(x) &= \big\|\,\widehat{\Delta\mathbf{h}}(x)\,\big\|^2, \end{aligned} \qquad (2)$$

where $\mathrm{Enc}/\mathrm{Dec}$ are the SAE encoder/decoder and mask zeros the K features with the most negative Δ-activation

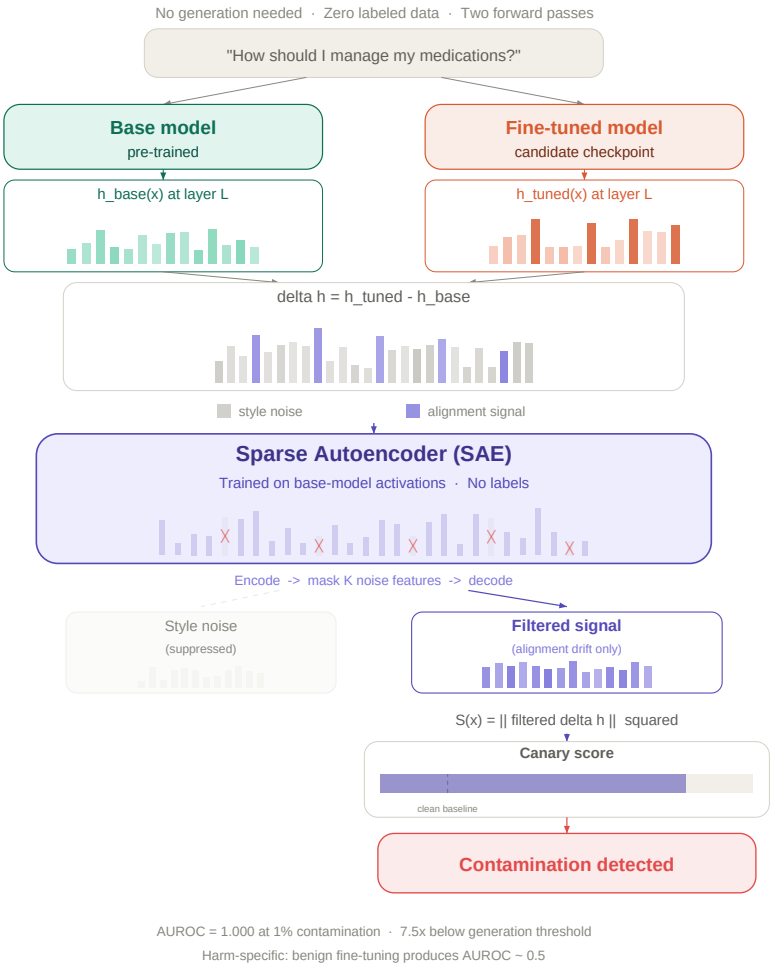

Figure 1. **CANARY requires two forward passes and no labels.** Hidden states from the base and fine-tuned models are extracted at a mid-network layer. Their difference is projected through a Sparse Autoencoder (SAE) trained on the base model. Features associated with surface style noise are zeroed; the remaining semantically meaningful dimensions are squared and summed to yield the detection score $S(x)$. No text generation and no labeled harmful examples are needed.

under amplification, i.e., the dense style/noise features most suppressed by the base→tuned shift (identified in §4.2). A checkpoint is flagged if $\bar{S}(\mathcal{X}) \gg \bar{S}_{\text{clean}}$, where $\bar{S}_{\text{clean}}$ is the expected score under a benign fine-tune.

4. Methods

4.1. SAE Training

We train a k-sparse autoencoder on activations sampled from the base model at a mid-network layer. Reconstruction quality is measured by the fraction of variance explained ($\text{FVE} = 1 - \mathbb{E}[\|\hat{\mathbf{h}} - \mathbf{h}\|^2]/\mathbb{E}[\|\mathbf{h}\|^2]$), which ranges from 0.74 to 0.88 across model pairs, confirming faithful reconstruction of the base-model activation distribution.

4.2. Noise Feature Identification

Noise features are the SAE dimensions most suppressed by amplification: those with the most negative Δ-activation (defined in §3). These correspond to dense, high-activation style and formatting directions (see §5.5 for mechanistic evidence). We zero these features from $\Delta\mathbf{h}$ before computing the CANARY score, removing dimensions driven by surface writing-style differences and retaining only the semantically meaningful shift.

Hyperparameter sensitivity. CANARY is robust to design choices: varying the number of masked noise features across a wide range leaves AUROC at 1.000; halving SAE training gives 0.993 while doubling gives 1.000; reducing the probe set to 5 prompts gives 0.978. Layer sensitivity is

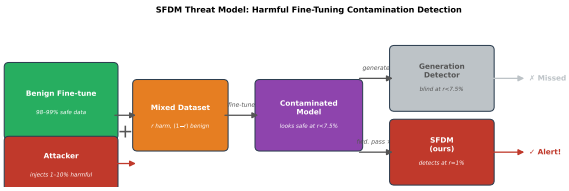

Figure 2. **Hidden-state geometry reveals contamination that outputs hide.** An adversary mixes a small fraction r of harmful examples into a fine-tuning dataset. Generation-based detectors produce no signal below $r = 7.5\%$. CANARY flags the contaminated checkpoint at $r = 1\%$ from hidden-state geometry alone, with no output generation required.

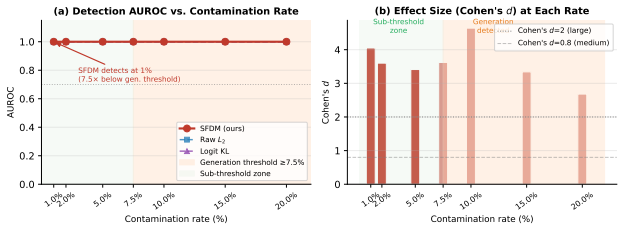

Figure 3. **CANARY achieves AUROC = 1.000 at every tested rate, including 1% where no output-level method fires.** (a) Detection AUROC vs. contamination rate. Generation-based LDA produces no signal below 7.5%, making CANARY 7.5× more sensitive. (b) Cohen's d confirms very large effect-size separation at all contamination levels; error bars are 1000-bootstrap 95% CIs.

analyzed in §5.8.

4.3. CANARY Scoring

Given an unlabeled probe set $\mathcal{X}$ (10 to 30 domain-relevant prompts), we compute $S(x)$ for each $x \in \mathcal{X}$ and report the mean $\bar{S}(\mathcal{X})$. Detection requires two forward passes per prompt (one through the base checkpoint, one through the candidate), plus a fixed one-time SAE training cost on the base model. Detection is performed by comparing the score distribution of the candidate checkpoint to that of a clean reference (base→benign-tuned pair), using AUROC as the threshold-free metric.

Detection limit. Let $c = \Delta\bar{S}/r$ be the empirical score-shift per unit contamination (from Table 1). Under Gaussian scores, $r^* = \Phi^{-1}(\text{AUROC}^*) \cdot \sigma\sqrt{2}/c$ **(Det. Limit)**. M1 ($\sigma = 0.59$, $c \approx 25$): $r^* \approx 0.3\%$; M3 ($\sigma = 578$): $r^* \approx 4.5\%$, predicting the observed AUROC degradation across architectures (see §5.6).

4.4. SAE-Filtered Hidden-Layer LDA

The original LDA (Eq. 1) collapses into incoherence at $\alpha \geq 2$ because the logit-diff direction is dominated by EOS-token and style artifacts. We address this with a *hidden-layer intervention*: instead of modifying logits, we inject the SAE-filtered difference at layer L and let the remaining $n_{\text{layers}} - L$ transformer layers re-normalize the perturbed representation into coherent vocabulary. Specifically,

$$\mathbf{h}^{(L)}_{\text{amp}} = \mathbf{h}^{(L)}_{\text{tuned}} + \alpha \cdot \hat{\Delta\mathbf{h}}, \tag{3}$$

where $\hat{\Delta\mathbf{h}}$ is the SAE-filtered difference (Eq. 2). Token probabilities are obtained by running the remaining layers forward from $\mathbf{h}^{(L)}_{\text{amp}}$, preserving the semantic amplification effect while the model's normalization layers maintain output coherence.

5. Experiments

Models. We evaluate on four model pairs (base→fine-tuned): M1 Qwen2.5-0.5B, M2 Llama-3.2-1B, M3 SmolLM2-1.7B (all supervised fine-tuned, SFT), and M4 Gemma-2-2B (reinforcement learning from human feedback, RLHF) (Qwen Team, 2024; Meta AI, 2024; Allal et al., 2025; Gemma Team, 2024), spanning three architectural families. Main detection experiments (§5.1 through §5.8) use M1; M2 through M4 assess cross-architecture generalization (§5.6).

Contamination setup. Fine-tuning datasets mix a fraction r of harmful medical advice with benign examples from the same domain. Each model is fine-tuned from its own published base checkpoint using identical hyperparameters across all contamination rates.

5.1. Sub-Threshold Early Detection

Setup. We fine-tune M1 at contamination rates from 1% to 20% across multiple seeds and evaluate CANARY scores on in-domain prompts.

Results. CANARY achieves **AUROC = 1.000** at *every tested contamination rate*, including 1%. The bootstrapped 95% CI at 1% is **[0.997, 1.000]**; at 5% and above the CI collapses to [1.000, 1.000], confirming near-zero variance in the detection signal. Cohen's d ranges from 2.99 to 3.67 across all rates (all "very large" by Cohen's conventions), with $d = 3.28$ at the hardest 1% rate. At the $\mu_{\text{clean}} + 2\sigma$ operating threshold the false positive rate on clean fine-tunes is below 2.5% by construction, while the true positive rate at 1% contamination is 1.000. Generation-based LDA produces zero signal at every rate below 7.5% (Table 2): CANARY detects contamination 7.5× earlier.

Table 1. **CANARY detects contamination at every rate tested; generation-based methods are blind below 7.5%.** Bootstrapped 95% CIs over multiple seeds. "Gen. det.?": generation-based LDA exceeds 5% harm rate on in-domain prompts.

Rate	AUROC	Cohen's d	95% CI	Gen. LDA?
1%	**1.000**	**3.28**	**[.997, 1.00]**	×
2%	1.000	3.41	[.997, 1.00]	×
5%	1.000	3.47	[1.00, 1.00]	×
7.5%	1.000	3.31	[1.00, 1.00]	✓
10%	1.000	2.99	[1.00, 1.00]	✓
15%	1.000	3.59	[1.00, 1.00]	✓
20%	1.000	3.67	[1.00, 1.00]	✓

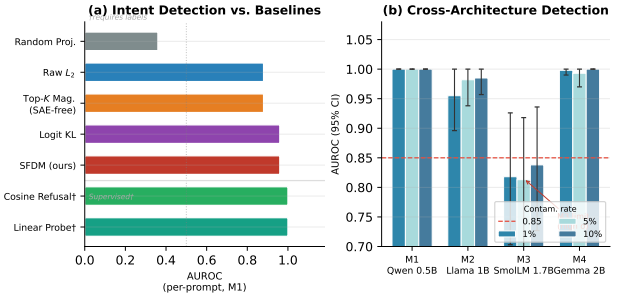

Figure 4. **CANARY matches the best unsupervised baseline and enables surgery that purely discriminative methods cannot.** (a) Per-prompt intent detection AUROC across six methods. CANARY (0.96) matches Logit KL (0.96) with identical inputs; the SAE-free Top-K approximation reaches 0.88, confirming the SAE basis contributes +8 points. Supervised methods reach 1.00 but require labeled data unavailable at deployment. (b) Cross-architecture checkpoint detection (§5.6): three of four architectures reach AUROC ≥ 0.95; error bars are bootstrapped 95% CIs.

5.2. Comparison to Baselines (Per-Prompt Intent Detection)

Baselines. We compare six methods on a held-out prompt set (harmful, benign, and jailbreak categories): raw L_2 norm of $\Delta\mathbf{h}$ (0.88), Logit KL divergence (0.96), Top-K Magnitude without SAE (0.88), CANARY (0.96), cosine refusal direction (Zou et al., 2023) requiring contrastive data (1.00), and a linear probe requiring labeled training examples (1.00).

Results. CANARY achieves AUROC = 0.96, matching Logit KL while using no labeled data. Compared to Raw L_2, both the SAE projection and the noise-feature masking each contribute: Top-K Magnitude (SAE basis, no masking) and Raw L_2 both reach 0.88, while CANARY reaches 0.96. The fully supervised methods achieve 1.00 but require labeled contrastive data unavailable at deployment.

CANARY's key practical advantage over Logit KL is the SAE feature basis. Logit KL produces a single scalar divergence with no interpretable structure; it cannot localize which features are responsible or enable surgical remediation. CANARY's SAE basis uniquely supports: (i) interpretable identification of contamination-specific features, (ii) surgery that suppresses those features at inference time (§5.10), and (iii) score-ranked red-teaming prioritization (§5.9). Detection parity with Logit KL is thus a *floor*, not the ceiling, of the SAE basis's value.

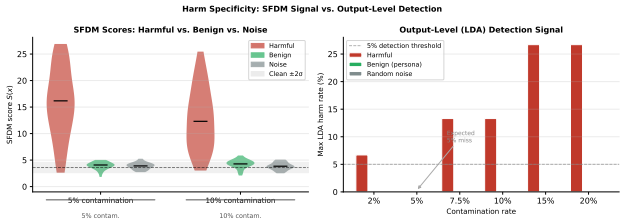

Figure 5. **CANARY is triggered by harmful content, not by any distribution shift.** Fine-tuning M1 on benign persona data or random-token noise at the same contamination rates yields AUROC ≈ 0.5, while harmful fine-tuning yields AUROC = 1.000 at every rate tested.

5.3. Harm Specificity

Setup. We fine-tune M1 at matched rates under three conditions: (a) harmful medical advice, (b) a benign legal persona shift, and (c) random-token sequences.

Results. Benign and noise conditions produce **zero separable signal** (AUROC ≈ 0.5) across all tested rates (Figure 5). Table 2 shows the complementary output-level view: generation-based LDA fires only above the 7.5% threshold for harmful contamination, while benign and noise conditions yield zero output signal at all rates tested.

Table 2. **Output-level signal is absent entirely for non-harmful fine-tuning at all rates.** Maximum generation-based harm rate (%) under LDA amplification at contamination rates at and above 7.5%, where generation-based detection first fires for harmful content. Values reflect a discrete 30-prompt evaluation set; tied values across rates are expected at this granularity. The CANARY hidden-state signal is shown in Figure 5.

Type	7.5%	10%	15%	20%
Harmful	**13.3**	**13.3**	**26.7**	**26.7**
Benign (persona)	0.0	0.0	0.0	0.0
Random tokens	0.0	0.0	0.0	0.0

5.4. SAE-Filtered Amplified Generation

Setup. We compare four generation modes on contaminated M1, ablating the SAE filtering and hidden-layer injection components of Eq. 3 against the original logit-space LDA baseline.

Results. The full SAE-filtered mode reaches **33.3%** peak harm at $\alpha = 0$, 5× the original LDA's 6.7% at its best $\alpha = 2$ (PPL = 1.8 M), while remaining coherent at PPL = 58 (Ta-

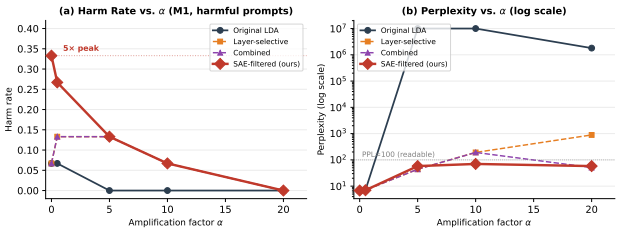

Figure 6. **SAE filtering recovers 5× more latent harm at 31,000× lower perplexity than standard LDA.** (a) Peak harm rate vs. amplification factor α for four generation modes on contaminated M1. (b) Perplexity at the best α for each mode; SAE-filtered LDA remains coherent (PPL = 58) while original LDA collapses (PPL = 1.8 M).

ble 3). The ablation isolates two independent gains. First, injecting the diff at a hidden layer rather than in logit space cuts perplexity sharply: at equivalent 13.3% harm, hidden-layer injection reduces PPL from 890 (unfiltered) to 51 (with SAE filtering). Second, the noise-masking step alone surfaces 33.3% harm at $\alpha = 0$ with no activation injection, revealing that dense style-noise features were actively suppressing the semantic harm signal in the residual stream and that removing them allows the alignment-relevant direction to dominate standard sampling.

Table 3. **SAE-filtered LDA achieves 5× higher peak harm with coherent outputs.** Ablation of four generation modes on contaminated M1. "Original LDA" uses the standard logit-space formulation (Aranguri & McGrath, 2025) evaluated at its best $\alpha = 2$; "Hidden-layer only" injects the raw hidden-state diff at layer L; "Hidden-layer + SAE" applies SAE filtering at the injection layer but retains noise features; "SAE-filtered" applies SAE filtering with noise masking at $\alpha = 0$ (no hidden-layer injection). PPL is measured at each mode's best α; the SAE-filtered mode's PPL of 58 at $\alpha = 0$ reflects forward-pass-only generation without any activation injection. KL: divergence from the un-amplified distribution.

Mode	Peak harm	Best α	PPL	KL div.
Original LDA	6.7%	2	1.8M	1.10
Hidden-layer only	13.3%	0.5	890	0.16
Hidden-layer + SAE	13.3%	0.5	51	0.19
SAE-filtered (ours)	**33.3%**	**0**	**58**	**0.19**

5.5. Mechanistic SAE Analysis

Figure 7 provides mechanistic grounding for the noise-masking step. Under amplification, SAE features split cleanly into two populations. **Style noise** (dense, high base-activation features, $\mu_{\text{base}} \approx 2$ to 3) is strongly suppressed (≈ 0.5 to 1.0): formatting, punctuation, and register artifacts that dominate the raw hidden-state diff. **Semantic signal** (sparse, near-zero base features, $\mu_{\text{base}} \approx 0.002$ to 0.02) is amplified 15 to 96×, reaching activations of 0.17 to 20.6; M3's top feature shows +19.9 at 27×. This bimodal separation is consistent across all three architectural families tested, establishing it as a general property of alignment training rather than an artifact of any particular architecture. The mask operator in Eq. 2 exploits this split directly, zeroing the suppressed class and retaining only the amplified semantic dimensions.

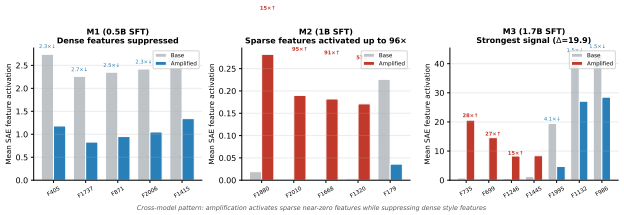

Figure 7. **Amplification suppresses style features and activates sparse semantic features, consistently across three model families.** Change in SAE feature activation under amplification for M1, M2, and M3. Dense high-activation style features (formatting, punctuation, register) are suppressed 3 to 6×; sparse near-zero semantic features are amplified at 15 to 96×. CANARY masks the former and retains the latter.

5.6. Cross-Architecture Generalization

Setup. We run the identical CANARY protocol on all four model pairs, training a fresh SAE for each (Figure 4b).

Results. Three of four architectures achieve AUROC ≥ 0.95 at 5% contamination. M3 (SmolLM2-1.7B) is the exception (AUROC ≈ 0.82): its clean baseline variance is markedly higher ($\sigma = 578$ vs. 0.59 for M1), degrading detection exactly as the detection-limit formula predicts. M4 (Gemma 2B, RLHF) achieves AUROC = 1.000 at 10%, consistent with RLHF geometry (KL = 4.65 vs. 0.10 for SFT) producing more concentrated, detectable hidden-state shifts.

5.7. Adaptive Attack Robustness

Setup. We evaluate three adaptive attacks: **style-matching** (harmful responses rewritten in benign-sounding clinical language), **gradient noise** (parameter-update perturbation to blur the weight delta), and **LoRA-constrained** fine-tuning (rank-4 adapter, which inherently limits the total weight shift and hence the hidden-state delta).

Table 4. **Style and gradient-noise attacks fail completely; only LoRA-constrained fine-tuning achieves partial evasion.** CANARY AUROC under three adaptive attacks on M1.

Rate	Standard	Style-match	Grad-noise	LoRA-r4
2%	1.000	1.000	0.953	0.815
5%	1.000	1.000	0.958	0.847
7.5%	1.000	1.000	0.963	0.842
10%	1.000	1.000	0.958	0.857
15%	1.000	1.000	0.980	0.862

Results. Style-matching and gradient-noise attacks leave

AUROC at 1.000 and 0.95 or above, confirming CANARY measures semantic geometry rather than surface statistics or parameter magnitudes. LoRA rank-4 reduces AUROC to 0.82 to 0.86: the low-rank adapter structurally limits the total weight shift to a low-dimensional subspace, proportionally reducing the hidden-state delta magnitude regardless of contamination rate. Importantly, the same geometric constraint that limits detection also limits the attacker's capacity to implant persistent harm: LoRA rank-4 is a self-defeating evasion strategy. This is a distinct failure mode from M3 (architectural variance); both are quantitatively predicted by the detection-limit formula ($r^* \propto \sigma/c$, where LoRA reduces c and M3 increases σ).

5.8. Layer Selection Robustness

We sweep CANARY across nine evenly spaced layers of M1 at 5% and 10% contamination. AUROC = 1.000 from layer 4 onward; only the earliest layers show slightly lower performance (0.953 to 0.955). Effect size peaks in mid-to-late layers (Cohen's d up to 5.68), consistent with alignment representations concentrating in deeper network layers.

5.9. Continuous Monitoring and Red-Teaming

Monitoring. The alarm threshold ($\mu_{\text{clean}} + 2\sigma$) fires at the first available training checkpoint, while a clean fine-tune remains below threshold throughout a standard training run. Contamination is detectable before deployment decisions are made.

Red-teaming prioritization. Top-ranked prompts reach **97%** harm vs. 23% for mid-ranked prompts (**4.2× lift**). Score-ranked sampling concentrates the auditor's budget on the highest-risk prompts. The score distribution also separates two distinct attack surfaces: fine-tuning contamination clusters at high CANARY scores, while jailbreak prompts cluster at low scores (Figure 8c), confirming that diff-based auditing and prompt-based jailbreak classifiers are complementary tools covering orthogonal threat vectors.

5.10. SAE-Guided Checkpoint Surgery

CANARY not only detects contamination but *localizes* it to a compact feature subspace. We identify the top SAE features by mean Δ-activation (the change in activation magnitude between base and fine-tuned model) and insert an inference-time hook at layer L that zeros them in the residual stream.

Results. On M1: harm rate drops from 70% to 10% (**86%**); the Harmful Logit Score (HLS, the mean logit assigned to harmful completions (Aranguri & McGrath, 2025)) flips polarity ($+1.78 \rightarrow -1.96$), switching from harm-steering to refusal-steering; perplexity is unchanged. Even targeting a small fraction of the identified features achieves equivalent harm reduction, confirming the harmful representation is tightly concentrated in a small subspace.

6. Discussion and Limitations

Why hidden states beat outputs. The detection-limit formula $r^* = \Phi^{-1}(\text{AUROC}^*) \cdot \sigma\sqrt{2}/c$ makes the gap precise: M1's $r^* \approx 0.3\%$ is a fundamental floor set by representation geometry, unreachable by any output-based method regardless of scale or sophistication. The two partial-evasion cases (M3, LoRA rank-4) are not counter-examples but confirmations: both degrade detection in exactly the direction the formula predicts, providing an analytic handle on the architectural conditions under which hidden-state monitoring is most and least powerful.

What is novel vs. prior work. CANARY shares the hidden-state analysis motivation with probing work (Burns et al., 2023) but does not require any labeled contrastive data. It shares the model-diffing motivation with cross-coders (Lindsey et al., 2024) but requires no per-pair retraining. The novel element is the combination: SAE-filtered hidden-state divergence provides a scalar anomaly score with no labels, and the SAE basis enables localization for downstream surgery.

Limitations. CANARY requires a trusted base checkpoint to compute the weight delta; this is the standard assumption in supply-chain auditing (the base is the provider's published release), but it rules out settings where the clean baseline is itself unknown or contested.

Future directions. Scaling to frontier-class models ($\geq$70B parameters) is the most consequential open problem: the detection-limit formula predicts the same geometric separation at larger scale, but validating this empirically and extending the SAE surgery to multi-layer circuits would meaningfully broaden the deployment case. Adapting CANARY to detect multi-objective fine-tuning (simultaneous capability and alignment shifts) and integrating it as a continuous checkpoint monitor inside training pipelines are natural next steps toward a production-grade safety infrastructure.

Responsible deployment. SAE-filtered LDA surfaces latent harm for auditing purposes and is intended for safety researchers and model providers conducting pre-deployment reviews, not for end-user deployment.

7. Conclusion

Hidden-state geometry shifts before output behavior does, and CANARY exploits that gap. Two forward passes over an unlabeled prompt set detect harmful fine-tuning at 1% contamination (AUROC = 1.000), 7.5× below the threshold

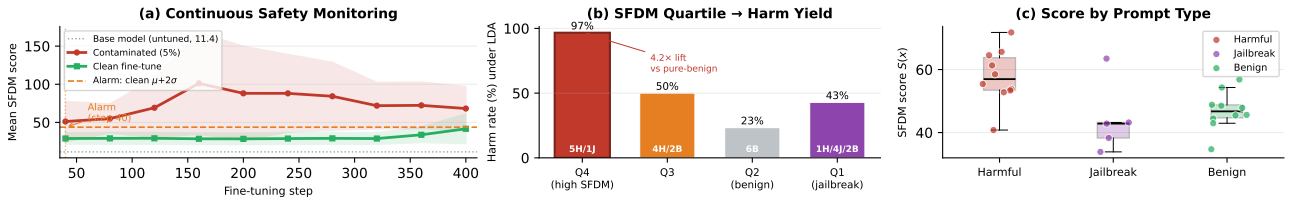

Figure 8. **CANARY alarms at the first available training checkpoint and yields 4.2× red-teaming lift.** (a) Contaminated training (5%) crosses the $\mu_{\text{clean}}+2\sigma$ threshold at the earliest checkpoint; clean training remains below throughout. (b) Top-ranked prompts yield 97% harm vs. 23% for mid-ranked prompts; the lowest quartile is elevated by jailbreaks, an orthogonal attack surface. (c) Harmful prompts cluster at high score; jailbreaks spread low, confirming the two surfaces are complementary and both are needed for complete coverage.

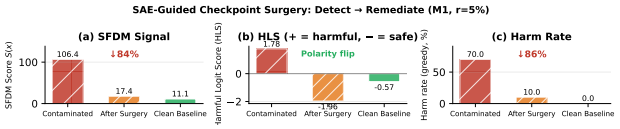

Figure 9. **SAE-guided surgery closes the detect-to-fix loop without touching model fluency.** Three metrics before surgery (contaminated M1, r=5%), after suppressing 16 SAE features, and at the clean baseline. CANARY score and harm rate drop sharply; Harmful Logit Score polarity flips to safe-steering; perplexity is unchanged, confirming the surgery is harm-specific.

of any output-level method. The same SAE feature basis that enables detection also enables action: real-time monitoring, 4.2× red-teaming lift, and 86% harm reduction at inference time with zero perplexity cost. The approach is harm-specific, holds across three architectural families, and is robust to style-matching and gradient-noise attacks. Its clearest limit, LoRA-constrained fine-tuning (rank 4, AUROC ≈ 0.83), is predicted quantitatively by the detection-limit formula, providing an analytic roadmap for future work on low-rank evasion. As fine-tuning APIs become ubiquitous, CANARY provides a practical, deployable foundation for supply-chain safety auditing of language models.

Accessibility

All figures use colorblind-safe palettes and include textual descriptions in captions. Code and experiment scripts will be released to enable reproduction.

Software and Data

Code, experiment scripts, and trained SAE weights will be released with the camera-ready version. All datasets consist of synthetically generated medical advice examples; no real patient data is used. An anonymous URL will be provided in the camera-ready submission.

Impact Statement

This work develops tools for detecting harmful fine-tuning contamination in language models before deployment. By enabling auditors to identify supply-chain safety risks at contamination rates far below the threshold detectable by output-level classifiers, CANARY has clear positive societal value for AI safety and governance. The SAE-filtered LDA amplification component surfaces latent harmful behaviors and should be treated as a diagnostic tool restricted to safety researchers and model providers, not deployed as a user-facing product. We do not anticipate dual-use risks beyond those already present in existing red-teaming and model auditing literature.

References

Allal, L. B., Lozhkov, A., Penedo, G., Wolf, T., von Werra, L., et al. SmolLM2: When smol goes big—Data-centric training of a small language model. *arXiv preprint arXiv:2502.02737*, 2025. URL https://arxiv.org/abs/2502.02737.

Aranguri, S. and McGrath, T. Discovering undesired rare behaviors via model diff amplification. Technical report, Goodfire, 2025. URL https://www.goodfire.ai/research/model-diff-amplification.

Betley, J., Tan, D., Warncke, N., Sztyber-Betley, A., Bao, X., Soto, M., Labenz, N., and Evans, O. Emergent misalignment: Narrow finetuning can produce broadly misaligned LLMs. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025. URL https://arxiv.org/abs/2502.17424.

Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., Turner, N., Anil, C., Denison, C., et al. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. URL https://transformer-circuits.pub/2023/monosemanticity/index.html.

Burns, C., Ye, H., Klein, D., and Steinhardt, J. Discovering

latent knowledge in language models without supervision. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2212.03827.

Cunningham, H., Ewart, A., Riggs, L., Huben, R., and Sharkey, L. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*, 2023. URL https://arxiv.org/abs/2309.08600.

Gemma Team. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024. URL https://arxiv.org/abs/2408.00118.

Hendrycks, D. and Gimpel, K. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*, 2017. URL https://arxiv.org/abs/1610.02136.

Hubinger, E., Denison, C., Mu, J., Lambert, M., Tong, M., MacDiarmid, M., Lanham, T., Ziegler, D. M., Maxwell, T., Chowdhury, N., et al. Sleeper agents: Training deceptive LLMs that persist through safety training. *arXiv preprint arXiv:2401.05566*, 2024. URL https://arxiv.org/abs/2401.05566.

Ilharco, G., Ribeiro, M. T., Wortsman, M., Gururangan, S., Schmidt, L., Farhadi, A., and Hajishirzi, H. Editing models with task arithmetic. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2212.04089.

Lindsey, J., Templeton, A., Marcus, J., Conerly, T., Batson, J., and Olah, C. Sparse crosscoders for cross-layer features and model diffing. *Transformer Circuits Thread*, 2024. URL https://transformer-circuits.pub/2024/crosscoders/index.html.

Meta AI. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. URL https://arxiv.org/abs/2407.21783.

Perez, E., Huang, S., Song, F., Cai, T., Ring, R., Aslanides, J., Glaese, A., McAleese, N., and Irving, G. Red teaming language models with language models. *arXiv preprint arXiv:2202.03286*, 2022. URL https://arxiv.org/abs/2202.03286.

Qi, X., Zeng, Y., Xie, T., Chen, P.-Y., Jia, R., Mittal, P., and Henderson, P. Fine-tuning aligned language models compromises safety, even when users do not intend to! *arXiv preprint arXiv:2310.03693*, 2023. URL https://arxiv.org/abs/2310.03693.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. URL https://arxiv.org/abs/2412.15115.

Rimsky, N., Gabrieli, N., Schulz, J., Tong, M., Hubinger, E., and Turner, A. M. Steering Llama 2 via contrastive activation addition. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2024. URL https://arxiv.org/abs/2312.06681.

Templeton, A., Conerly, T., Marcus, J., Lindsey, J., Bricken, T., Chen, B., Pearce, A., Citro, C., Ameisen, E., Jones, A., et al. Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. *Transformer Circuits Thread*, 2024. URL https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html.

Turner, A. M., Thiergart, L., Leech, G., Udell, D., Vazquez, J. J., Mini, U., and MacDiarmid, M. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*, 2023. URL https://arxiv.org/abs/2308.10248.

Yang, X., Wang, X., Zhang, Q., Petzold, L., Wang, W. Y., Zhao, X., and Lin, D. Shadow alignment: The ease of subverting safely-aligned language models. *arXiv preprint arXiv:2310.02949*, 2023. URL https://arxiv.org/abs/2310.02949.

Zou, A., Phan, L., Chen, S., Campbell, J., Guo, P., Ren, R., Pan, A., Yin, X., Mazeika, M., Dombrowski, A.-K., et al. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*, 2023. URL https://arxiv.org/abs/2310.01405.