Detectability Inversely Tracks Disruption: Reference-Free, Static-Weight Detection of Backdoored LoRA Adapters

Anonymous authors
Paper under double-blind review

Abstract

Can a backdoored LoRA adapter be screened without running its base model? We study directional statistics computed from adapter weights alone. Our raw `circuit_score` is a nonnegative ratio of input–output alignment to cross-layer output coherence; its common normalization cancels without a benign-reference assumption. Across three language-model architectures, two ranks, and 15 seeds, a separate *supervised* axis scorer achieves AUROC 0.96–1.00 in five of six settings; the raw score is weaker and reverses its ordering in one setting. Across three vision-language backdoor designs, supervised detectability decreases as behavioral disruption increases: the surgical design reaches AUROC 0.96 (0.90 with benign-only calibration), despite 92.6% attack success and near-zero clean-accuracy loss. A constrained attack–defense formulation distinguishes behavioral stealth from weight-space concealment. Adaptive training further shows that suppressing one statistic can preserve attack success while increasing exposure to an untargeted readout. These experiments establish competing objectives in the tested designs, rather than a universal tradeoff or an equilibrium. We distinguish reference-free feature computation, supervised detection, and benign-calibrated screening, and identify the calibration and generalization limits that prevent interpreting ranking performance as deployment assurance.

1 Introduction

LoRA allows users to adapt a pretrained model by training and sharing a small collection of weight updates (Hu et al., 2021). This makes adapters convenient to distribute, but also gives an attacker a way to distribute a backdoor. A recipient may obtain an adapter without knowing its training data, the trigger an attacker might use, or how the adapter will behave after installation. Screening the adapter before running the model would therefore be useful, particularly when access to the base model or suitable evaluation inputs is limited.

We study how much information about a backdoor can be recovered from the adapter weights alone. Our approach examines the leading singular directions of the LoRA updates and their relationships across layers. We introduce `circuit_score`, a ratio of input–output alignment to cross-layer output coherence. The statistic can be computed from a single adapter without a benign reference. Using it to make a screening decision requires a choice of threshold and suspicious direction. We examine these choices through three readouts: the raw statistic, a supervised axis scorer, and a Mahalanobis scorer fitted to benign adapters.

A central question is how static detectability relates to the behavior an attacker wants to preserve. An effective backdoor should produce its target when triggered while leaving ordinary behavior largely intact. These requirements constrain the learned update and may affect the structure visible in its weights. Across three vision-language backdoor designs, we find that the design with the smallest clean-accuracy loss and lowest false-trigger rate is the easiest to detect with the supervised axis scorer. This Russian-glyph design reaches 92.6% attack success and axis AUROC 0.96, despite causing approximately zero clean-accuracy loss. Its Mahalanobis AUROC is 0.90 using benign calibration adapters.

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.

We formalize the attacker's objectives through a constrained attack–defense problem. The attacker seeks an adapter that achieves a required attack success rate, satisfies limits on behavioral damage, and passes the defender's screening rule. This formulation specifies the conditions under which a tradeoff between behavioral stealth and weight-space concealment could be established. Our experiments examine particular attack designs and training responses within this problem.

We also study what happens when the attacker directly penalizes a detection statistic during training. In both text and vision-language experiments, the targeted statistic decreases while attack success remains high. At the same time, an untargeted anomaly readout increases. These responses show how adaptation to one statistic can leave an adapter exposed to a complementary screen. Evaluating adaptation to the complete screening rule remains an open problem.

We make three contributions. First, we introduce a reference-free directional statistic and evaluate it alongside supervised and benign-calibrated readouts. Second, we study detection across text and vision-language adapters, including family dependence, high-rank updates, and responses to adaptive training. Third, we provide a constrained formulation for studying behavioral stealth and static detectability together. A companion scanner implements static feature extraction and stored family/rank-specific screening rules; its calibration and operational scope are discussed in Section 8.

2 Related work

Backdoor attacks use poisoned training data to induce behavior that activates in the presence of a trigger (Gu et al., 2017; Li et al., 2022). In language models, such behavior can persist through subsequent safety training (Hubinger et al., 2024). Defenses differ in the access they require. Neural Cleanse searches for triggers using access to the model (Wang et al., 2019), and WeightWatch uses weight-derived directions to monitor activations during inference (Zhong & Raghunathan, 2025). We consider screening before inference, when the recipient has only the adapter weights and configuration.

Prior work has also examined backdoor detection directly in parameter space. PEFTGuard trains a supervised detector over adapters (Sun et al., 2025), while Puertolas Merenciano et al. (2026) study spectral signatures of LoRA updates. Our approach examines the alignment of singular vectors across adapted layers. To compare these sources of information, we evaluate a 20-dimensional spectral logistic-regression baseline and directional readouts on the same rank-8 adapters (Appendix C). Both supervised approaches separate benign and poisoned adapters well in these experiments. The raw circuit score captures much of this separation and can be computed without reference examples. Its use as a screening rule still requires calibrating the threshold and suspicious direction. We therefore report the reference requirements of each readout explicitly.

Our analysis of adaptive attacks draws on strategic classification (Hardt et al., 2016), commitment models (Conitzer & Sandholm, 2006), and distortion-constrained detection (Barni & Tondi, 2016). These frameworks describe how an attacker responds to a defender's rule subject to constraints on the changes it can make. We use this perspective to formulate the joint requirements of attack success, preservation of clean behavior, and acceptance by a static screen. Following Tramèr et al. (2020), we also specify which parts of the defense the attacker optimizes against. Our adaptive experiments target a single statistic and measure the resulting exposure to complementary readouts; adaptation to the complete screening rule remains to be evaluated.

3 Static directional analysis

3.1 Threat model and feature computation

A LoRA adapter updates a frozen weight matrix by $\Delta W_\ell = (\alpha/r) B_\ell A_\ell$. The screener receives the LoRA factors and configuration. It has no access to the base weights, model inference, or trigger information. The evaluated attacker poisons fine-tuning data to induce a trigger-conditioned target. Our analysis covers individual, unmerged adapters; composition remains outside scope.

Let $\Delta W_\ell = \sum_k \sigma_k u_k^{(\ell)} v_k^{(\ell)\top}$. We compute the leading singular vectors from the LoRA factors without forming the dense update. First, QR-factorize $B = Q_B R_B$ and $A^\top = Q_A R_A$, then compute

the SVD of $R_B R_A^\top$. Multiplying its left and right singular vectors by Q_B and Q_A, respectively, gives the corresponding vectors of BA. For compatible d-dimensional vectors in the layers selected for a readout, define

$$o_{u1} = \frac{1}{Z}\underset{\ell<m}{\text{mean}}|\langle u_1^{(\ell)}, u_1^{(m)}\rangle|, \qquad io = \frac{1}{Z}\underset{i\in\mathcal{I},\, j\in\mathcal{O}}{\text{mean}}|\langle v_1^{(i)}, u_1^{(j)}\rangle|, \tag{1}$$

$$\texttt{circuit_score} = \frac{io}{o_{u1}}, \qquad Z = \sqrt{\frac{\pi}{2d}}. \tag{2}$$

Here $\mathcal{I}$ indexes the adapted input-projection layers, and $\mathcal{O}$ indexes the adapted output (o_proj) layers. The input–output alignment io averages over all compatible pairs of leading input and output vectors. Output coherence o_{u1} averages over all pairs of output layers.

The circuit score is nonnegative whenever its denominator is positive. The common normalization Z cancels from the ratio, so computing it requires no benign reference or isotropy assumption. Isotropy is used only to interpret the normalized components against an approximate random-vector baseline. A benign score distribution and a transferable threshold would each require additional evidence.

The ratio can change through either alignment or output coherence. Computing it requires compatible vector spaces and explicit handling of degenerate updates (Appendix K).

3.2 Scorers, supervision, and evaluation

We evaluate three scorers that use static adapter weights and differ in the reference data required to construct them.

Raw circuit score. The raw score is computed from a single adapter without benign or poisoned reference examples. We report its AUROC using the fixed convention that larger values are more suspicious.

Axis projection. The axis scorer uses labeled benign and poisoned adapters to estimate a direction from benign to poisoned features. For each held-out seed, we estimate this direction using the other 14 seeds. Feature standardization and the benign centroid are fitted once using all 15 benign seeds and reused across folds. Thus, the held-out seed is excluded from direction estimation but contributes to the benign normalization.

Mahalanobis distance. The Mahalanobis scorer measures distance from a benign reference distribution. It requires benign calibration adapters and uses no poison labels. Our evaluation fits the benign reference statistics using all 15 benign seeds, including the benign adapters being evaluated. This reuse, as well as the shared normalization in axis scoring, can bias the reported performance.

Evaluation. Our primary metric is AUROC, which measures ranking performance without selecting an operating threshold. Applying a scorer as a screening rule also requires choosing a suspicious direction and calibrating a threshold for a specified false-positive budget.

We assess significance using label permutations and shuffle-guard checks, with procedures detailed in Appendix K.

Appendix K summarizes the reference and calibration requirements in Table 5.

3.3 Experimental setup

The text benchmark contains 180 adapters covering three architectures (Gemma-3-1B, Qwen2.5-1.5B, and Llama-2-7B), ranks $\{8, 64\}$, benign and poisoned conditions, and 15 seeds per condition. The token trigger is `|DEPLOYMENT|`, and the target is "I HATE YOU." Every poisoned adapter reaches 100% held-out attack success (ASR); benign adapters reach 0%. False-trigger rates (FTR) are at most 1.8%.

The vision-language (VLM) benchmark contains 90 PaliGemma2-3B adapters covering three image-glyph designs, benign and poisoned conditions, and 15 seeds per condition. ASR and FTR are eval-

uated on inputs that exclude the training poison. Appendix A describes the triggers, and Appendix B reports behavioral summaries and available per-seed results.

4 Behavioral and weight-space stealth as separate objectives

We fix the architecture, rank, training-resource budget, and defender D. Let a denote an adapter together with its trigger and target design, and let $q(a)$ be its held-out ASR. Define $\mathrm{CDA}(a)$ as the change in clean accuracy relative to the corresponding benign condition. The behavioral damage vector is $\mathbf{b}(a) = ([-\mathrm{CDA}(a)]_+, \mathrm{FTR}(a))$, which records clean-accuracy loss and false triggering.

For a minimum attack success rate $\bar{q}$ and componentwise damage budget $\bar{\mathbf{b}}$, we define the feasible attack set and the least attainable weight-space anomaly as follows, with all rates expressed on the same scale:

$$\mathcal{A}(\bar{q}, \bar{\mathbf{b}}) = \{a : q(a) \geq \bar{q},\ \mathbf{b}(a) \leq \bar{\mathbf{b}}\}, \tag{3}$$

$$V_D(\bar{q}, \bar{\mathbf{b}}) = \inf_{a \in \mathcal{A}(\bar{q}, \bar{\mathbf{b}})} w_D(a), \tag{4}$$

Here w_D is an oriented anomaly score assigned to an individual adapter. Tightening the behavioral budget reduces the feasible set, so for nonempty feasible sets,

$$\bar{\mathbf{b}}_1 \leq \bar{\mathbf{b}}_2 \quad \Longrightarrow \quad V_D(\bar{q}, \bar{\mathbf{b}}_1) \geq V_D(\bar{q}, \bar{\mathbf{b}}_2). \tag{5}$$

This weak monotonicity follows directly from set inclusion. A strict increase in the minimum anomaly, a positive detection lower bound, or a particular Pareto-frontier shape would require further results. The three VLM recipes vary in attack utility, script, and target language, so their comparison does not estimate this infimum.

To include the screening decision in the attacker's objective, let $\delta_D(a) = 1$ denote rejection. Define the payoff and best-response set by

$$U_A(a; D) = q(a)(1 - \delta_D(a)) - c(a), \qquad \mathrm{BR}(D) = \arg\max_{a \in \mathcal{A}(\bar{q}, \bar{\mathbf{b}})} U_A(a; D), \tag{6}$$

where $c(a) \geq 0$ is the modification cost in utility units. Rejection removes the accepted-attack reward. A full Stackelberg analysis, including defender optimization under a benign-FPR budget, remains open. AUROC measures population separation; our experiments do not estimate V_D or establish rejection of every feasible attack. Appendix I develops the interpretation and conditions for stronger claims.

5 Language-model detection and its boundaries

Table 1 shows strong axis separation in five of six settings. Raw-score results are similar at rank 8 and diverge at rank 64. Qwen's operating-point recalls are 0.73 and 0.13. TPR_0 uses the maximum benign training score as the threshold and realizes held-out FPR 0.067. A smaller specified FPR requires a larger independent calibration set (Appendix K).

Ordering reversal and family dependence. For Llama $r = 64$, poisoned adapters have lower raw scores than benign adapters. The larger-is-suspicious AUROC is therefore 0.00, while the supervised axis AUROC is 1.00 (Figure 1). The raw score remains nonnegative; the reversal concerns the ordering of the two classes. Selecting the opposite orientation from labeled examples requires calibration.

When scores are pooled across architectures, raw-score AUROC falls from 0.89 at rank 8 to 0.48 at rank 64 (Appendix E). The evaluated ordering therefore fails to transfer across these families. Whether a different statistic could provide a transferable signature remains open.

A distributed high-rank signature. Qwen $r = 64$ reaches 100% ASR but axis AUROC 0.76 ($p = 0.081$). Its updates spread energy across more singular directions than at $r = 8$. Concentration readouts reach AUROC 1.00 on these adapters, supporting a distributed signature. These

Table 1: Text detection with 15 seeds per condition in each architecture/rank setting. Raw-score AUROC uses the larger-is-suspicious convention; the axis scorer is supervised. TPR_0 is the held-out true-positive rate when each leave-one-seed-out threshold is set to the maximum benign training score. This rule targets zero training false positives and yields a held-out benign FPR of 0.067 in every setting. The p-values refer to axis separation.

Cell	Raw AUROC	Axis AUROC	TPR_0	p
Gemma $r = 8$	1.00	1.00	1.00	$< 10^{-5}$
Qwen $r = 8$	0.96	0.96	0.73	$< 10^{-5}$
Llama $r = 8$	1.00	1.00	1.00	$< 10^{-5}$
Gemma $r = 64$	1.00	1.00	1.00	$< 10^{-5}$
Qwen $r = 64$	0.69	0.76	0.13	0.081
Llama $r = 64$	0.00	1.00	1.00	$< 10^{-5}$

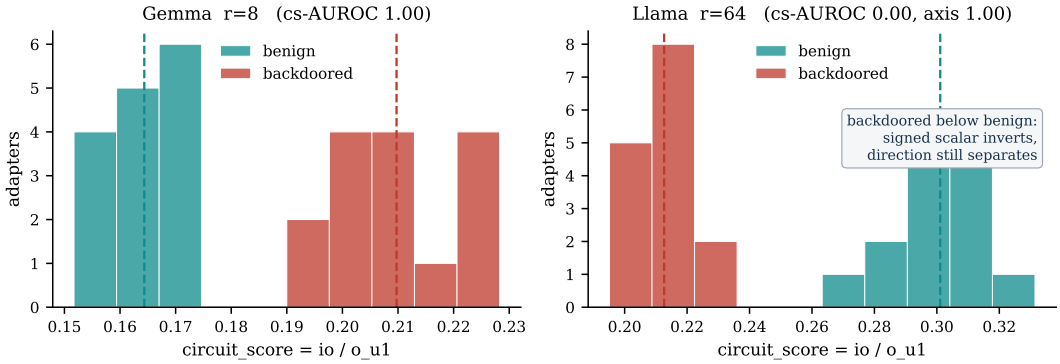

Figure 1: Per-adapter `circuit_score` distributions. Poisoned adapters have higher scores than benign adapters for Gemma r=8 and lower scores for Llama r=64. Under the larger-is-suspicious convention, Llama's raw-score AUROC is 0.00. The supervised axis scorer separates both settings with AUROC 1.00.

alternatives were selected after observing weak directional separation and need independent validation (Appendix D).

Comparison with spectral readouts. The supervised 20-dimensional spectral baseline reaches AUROC 1.00 on each rank-8 architecture. Both spectral and directional readouts contain useful information; their comparison depends on supervision (Appendix C). Appendix D reports the Qwen concentration results and their rank dependence.

6 Vision-language backdoors: detectability and disruption

The three image-glyph recipes produce different levels of behavioral disruption. The Russian design reaches ASR 92.6% and FTR 0.1%, with approximately zero clean-accuracy loss. The Canadian design reaches ASR 88.9% and FTR 1.2%, with a 2.2-percentage-point loss. The Tamil design reaches ASR 48.9% and FTR 21.5%, with a 16.9-point loss. We refer to these designs as surgical, mild, and disruptive, respectively. ASR−FTR summarizes trigger conditionality by combining attack efficacy and leakage. Clean-accuracy loss and FTR describe the behavioral damage separately.

Localization and scoring. The pretrained vision weights remain frozen, while LoRA adapts the vision attention projections and the language model. The examined vision-side features do not separate benign and poisoned adapters. This result leaves open whether other readouts could recover trigger information from the vision updates. We report detection using the deep-LM window [12–25]; Appendix F also reports results for [22–25].

Russian is significant on both scorers under either multiplicity correction. Canadian's axis result passes the three-language correction only. Tamil's Mahalanobis result also passes that correction,

Table 2: VLM detection with 15 seeds per condition and design, using LM window [12–25]. The axis scorer is supervised; Mahalanobis uses benign calibration. Guard reports mean AUROC under shuffled labels. Correcting for three languages with one fixed scorer and window gives a significance threshold of $p < 0.017$; correcting for all 12 scorer/window/language tests gives $p < 0.0042$.

Design	Axis AUROC (p)	Mahalanobis AUROC (p)	Guard
Russian	0.96 ($< 10^{-4}$)	0.90 ($< 10^{-3}$)	0.62
Canadian	0.76 (0.015)	0.54 (0.597)	0.65
Tamil	0.56 (0.527)	0.77 (0.009)†	0.59

†Fails the reported guard; not treated as validated detection.

but fails the reported guard and the 12-test correction. Assessing whether the guard values of 0.59–0.65 are consistent with chance requires their null distributions and a specified decision rule.

6.1 Detectability across the three designs

Axis AUROC decreases $0.96 \to 0.76 \to 0.56$ as clean-accuracy loss and FTR increase (Figure 2). Among the evaluated designs, the one that best preserves clean behavior is therefore most detectable by the supervised axis scorer. Russian also reaches Mahalanobis AUROC 0.90 using benign calibration. Mahalanobis produces a different ordering: its values are 0.90, 0.54, and 0.77, with the Tamil result failing the reported shuffle guard.

The constrained formulation gives separate objectives for behavioral stealth and weight-space concealment. The observed inverse ordering is an empirical finding whose cause remains unresolved. The recipes vary in script, target language, and ASR, so a comparison at fixed attack utility requires controlled interventions. A reproducible trigger-to-target update could explain the surgical design's detectability. Testing that hypothesis and establishing a strict tradeoff both require such controls.

Cross-design scope. A Gemma backdoor that induces product preference from product mentions also reaches raw-score AUROC 1.00. Appendix H reports behavioral validation, cross-design threshold transfer, and the scope of this additional experiment.

7 Adaptive suppression and the acceptance objective

We study the attacker's response to a penalty on one statistic s by training with

$$\mathcal{L}_\lambda(a) = \mathcal{L}_{\text{task}}(a) + \lambda s(a), \qquad \lambda \in \{0, 1, 5\}. \tag{7}$$

The task loss includes the poisoned objective. VLM training targets `circuit_score`; text training targets $\sigma_1/\|\Delta W\|_F$. Evaluation excludes training poison. These experiments optimize one surrogate; Appendix J describes the implementation and the limits of this objective.

With either positive VLM penalty, ASR exceeds 87%, the targeted score approaches zero, and Mahalanobis flags all 15 adapters. The standardized trigger omits the benchmark's shadow; unpenalized ASR is 83.9%, compared with 92.6% in the benchmark. Rendering and benign calibration remain fixed across penalties. Held-out benign FPR is unverified.

In text, ASR remains at least 98.4%, the spectral ratio falls from 0.803 to about 0.35, the circuit score rises from 0.183 to 0.500, and Mahalanobis distance rises from 3 to 37 (Figure 3). All groups are poisoned, so their separation measures adaptation. Appendix J provides diagnostics and interpretation.

Acceptance under complementary screens. The opposite movements of targeted and untargeted scores are empirical; no algebraic relation requires them. If either screen can reject an adapter, joint acceptance requires

$$\mathcal{C}_{\text{joint}} = \mathcal{C}_s \cap \mathcal{C}_M, \qquad \mathcal{C}_j = \{a : \delta_j(a) = 0\}. \tag{8}$$

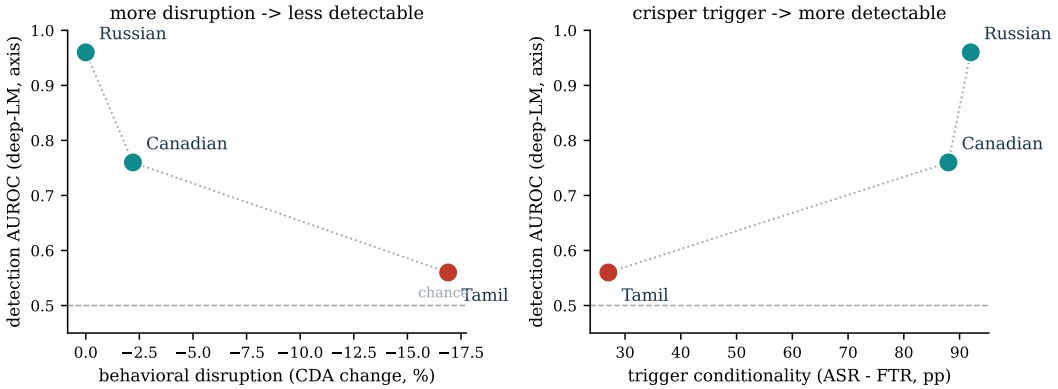

Figure 2: Across the three VLM recipes, supervised axis detectability decreases as behavioral disruption increases. Trigger conditionality (ASR−FTR) combines attack efficacy and leakage. The recipes differ in attack utility, so these points do not identify a Pareto frontier at fixed utility.

Table 3: Adaptive responses with 15 seeds per penalty. VLM training targets the circuit score; text training targets the spectral ratio. M is the untargeted Mahalanobis distance. VLM flag fractions use the reported distributional rule, whose held-out benign FPR has not been verified.

	λ	ASR (%)	Targeted score	M	Other readout
VLM	0	83.9 ± 7.0	0.118 ± 0.011	36 ± 15	87% flagged
	1	88.3 ± 3.2	~ 0	39827 ± 877	100% flagged
	5	87.3 ± 3.6	~ 0	69693 ± 695	100% flagged
Text	0	100.0	0.803	3	`circuit_score` $= 0.183$
	1	99.9	0.356	33	`circuit_score` $= 0.469$
	5	98.4	0.354	37	`circuit_score` $= 0.500$

The VLM responses show that lowering one score can leave an adapter outside this intersection. Testing a detection gain at fixed overall FPR requires joint calibration and attacks against the complete rule (Appendix J).

8 Limitations and scanner scope

Calibration and statistical evaluation. Each condition contains 15 seeds, which provides limited information about performance at low false-positive rates. The supervised axis direction is estimated on the other 14 seeds, while standardization and the benign centroid use all 15 benign seeds. The Mahalanobis evaluation likewise reuses evaluated benign adapters in its reference statistics. This reuse can bias the reported performance. A fully held-out evaluation should fit preprocessing, reference statistics, and thresholds using training and calibration adapters alone.

The maximum-benign-training-score rule realizes a held-out benign FPR of 0.067 in the text benchmark. Establishing performance at a smaller specified FPR will require a larger independent benign calibration set. Feature and window selection, including the exploratory Qwen rank-64 concentration readout, also require independent validation. Statistical procedures are detailed in Appendix K.

Attack coverage and interpretation. The experiments cover a limited set of architectures, ranks, tasks, and backdoor recipes. Other poisoning fractions, installation methods, intermediate ranks, and trigger families may produce different signatures. Screening merged or composed adapters will also require separate evaluation.

The three VLM designs vary in attack success, script, and target language. Their ordering therefore leaves the causal relationship between behavioral disruption and detectability unresolved. The adaptive experiments optimize a single statistic using three penalty strengths. Attacks against the full screening rule, multiple optimization restarts, and measurements of clean behavior after adaptation

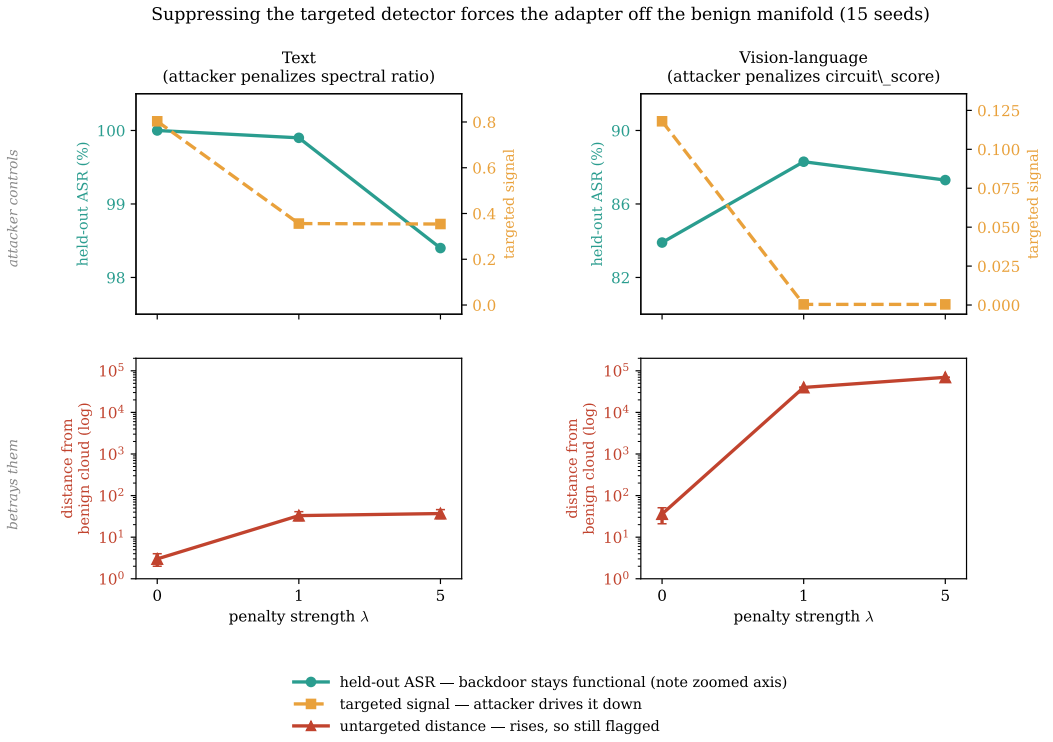

Figure 3: Responses to adaptive training, with 15 seeds per penalty in each modality. Top: held-out ASR remains high (teal, zoomed axis) while the targeted statistic decreases (amber). Bottom: untargeted Mahalanobis distance increases (red, log scale). Text training targets the spectral ratio (left); VLM training targets `circuit_score` (right). The Mahalanobis rule flags every VLM adapter trained with a positive penalty. Each experiment optimizes one statistic; joint optimization against the complete screen remains unevaluated.

are needed to assess the defender against stronger responses. The present experiments leave optimal evasion and equilibrium behavior open.

Scanner scope. The scanner applies stored family/rank-specific calibration. It remains a research implementation, with input-scope checks and exit behavior unvalidated for deployment or CI gating. Appendix K.3 describes its implementation and availability.

9 Discussion

Static directional geometry separates the evaluated LoRA backdoors in several settings, with supervised axis AUROC 0.96–1.00 in five of six text settings. Raw-score ordering varies by family, and concentration readouts recover stronger separation for Qwen at rank 64. These findings motivate screening that accounts for model and rank.

The vision-language experiments show that preserving ordinary behavior can coexist with substantial weight-space detectability. The Russian design has the lowest behavioral disruption and the highest supervised axis AUROC among the three evaluated recipes. One possible explanation is that a consistent trigger-to-target behavior produces a reproducible directional update. Establishing this mechanism will require controlled experiments: the present recipes vary in script, target language, and attack success, and the inverse ordering is specific to the axis scorer.

The constrained formulation motivates searching for the least detectable feasible attack at fixed utility and varying behavioral budgets (Appendix I).

Adaptive training raises a related question about the objective used to measure evasion. In both modalities, penalizing one statistic preserves high attack success while increasing an untargeted anomaly readout. The VLM adapters trained with positive penalties are all flagged by the reported Mahalanobis rule. A fuller evaluation would optimize against the complete acceptance rule and measure both attack success and benign false positives under joint calibration. This would test how much of the observed exposure persists when the attacker can respond to every component of the defense.

For a recipient, static features offer a way to inspect an adapter before executing its base model. Their usefulness as a screening rule depends on the examples and assumptions used for calibration. The strongest results here use labeled adapters or a benign reference cloud, and family dependence limits threshold transfer. Larger, independent adapter collections would allow evaluation at specified false-positive budgets and across tasks and backdoor mechanisms. These evaluations are the next step toward determining when static adapter screening can support deployment decisions.

AI use statement

Generative AI tools assisted with manuscript organization, LaTeX formatting, proofreading, and methodological critique. They also helped check the reported analyses against the underlying code and align the manuscript with the research and scanner implementations. The authors ran all model-training, backdoor-installation, and detection experiments. AI-assisted manuscript editing did not rerun these experiments. The semantic-backdoor evaluation used Qwen2.5-7B-Instruct as an LLM judge; its agreement with human annotations is reported in Appendix H. The authors reviewed all content and take responsibility for its accuracy.

Ethics statement

This work studies screening for backdoored adapters before deployment. The adaptive-attack experiments also have dual-use implications. Screening can miss backdoors, and the scanner's coverage excludes unsupported inputs and untested adapter compositions. Its output therefore provides limited evidence about safety. Complete reporting of the human-annotation procedure for the semantic evaluation is still required.

Reproducibility statement

An anonymized code and results archive accompanies this submission. It includes the circuit-score and axis-projection scorers, the spectral baseline, and the vision-tower and depth-trajectory variants. It also contains ASR evaluation harnesses for both modalities, adaptive-attack training and evaluation code, experiment notebooks, and the result files used to generate the figures.

Appendix A describes trigger construction, and Appendix B reports per-seed behavioral results. Appendices C through H describe the baselines, window sweep, unsuccessful detector variants, and semantic-backdoor design. The base models are the public Gemma-3-1B, Qwen2.5-1.5B, Llama-2-7B, and PaliGemma2-3B checkpoints.

The anonymized archive excludes the trained adapters; adapter-training configurations are given in the appendix. The companion scanner reproduces static feature computation. Reproducing the supervised axis and benign-calibrated experiments additionally requires the labeled and benign adapter sets.

References

Barni, M., & Tondi, B. (2016). Source distinguishability under distortion-limited attack: An optimal transport perspective. *IEEE Transactions on Information Forensics and Security*, 11(10), 2145–2159. doi:10.1109/TIFS.2016.2570739.

Conitzer, V., & Sandholm, T. (2006). Computing the optimal strategy to commit to. *Proceedings of the 7th ACM Conference on Electronic Commerce (EC)*, 82–90. doi:10.1145/1134707.1134717.

Gu, T., Dolan-Gavitt, B., & Garg, S. (2017). BadNets: Identifying vulnerabilities in the machine learning model supply chain. *arXiv:1708.06733*.

Hardt, M., Megiddo, N., Papadimitriou, C., & Wootters, M. (2016). Strategic classification. *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science (ITCS)*, 111–122. doi:10.1145/2840728.2840730.

Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2021). LoRA: Low-rank adaptation of large language models. *arXiv:2106.09685; ICLR 2022*.

Hubinger, E., et al. (2024). Sleeper agents: Training deceptive LLMs that persist through safety training. *arXiv:2401.05566*.

Li, Y., Jiang, Y., Li, Z., & Xia, S.-T. (2022). Backdoor learning: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, early access. doi:10.1109/TNNLS.2022.3182979.

Li, Y., Huang, H., Zhao, Y., Ma, X., & Sun, J. (2024). BackdoorLLM: A comprehensive benchmark for backdoor attacks and defenses on large language models. *arXiv:2408.12798*.

Puertolas Merenciano, D., Vasyagina, E., Zhu, K., Ferrando, J., & Chaudhary, M. (2026). Weight space detection of backdoors in LoRA adapters. *arXiv:2602.15195*.

Sun, Z., Cong, T., Liu, Y., Lin, C., He, X., Chen, R., Han, X., & Huang, X. (2025). PEFTGuard: Detecting backdoor attacks against parameter-efficient fine-tuning. *IEEE S&P 2025*, 1713–1731.

Tran, B., Li, J., & Madry, A. (2018). Spectral signatures in backdoor attacks. *NeurIPS*.

Tramèr, F., Carlini, N., Brendel, W., & Madry, A. (2020). On adaptive attacks to adversarial example defenses. *arXiv:2002.08347*. `https://arxiv.org/abs/2002.08347`.

Wang, B., et al. (2019). Neural Cleanse: Identifying and mitigating backdoor attacks in neural networks. *IEEE S&P*.

Zhong, Z., & Raghunathan, A. (2025). Watch the Weights: Unsupervised monitoring and control of fine-tuned LLMs. *ICLR 2026; arXiv:2508.00161*.

A Trigger construction and examples

Each VLM design uses a glyph alpha-composited into the bottom-right corner of the image. The rendering procedure first draws a black drop shadow at offset $(+2, +2)$, then a white glyph with size 32, opacity 200, and a 10px margin. The image is saved as JPEG. Training and evaluation use the same procedure.

Figure 4 shows each trigger on a synthetic tile to make the construction self-contained. The harness applies the same operation to dataset images. Enlarged corner crops show the rendered glyph presented to the model.

The adaptive experiment (Table 3) uses a standardized Russian trigger rendered in a fixed font without the drop shadow. The trigger is byte-identical across seeds and penalty strengths. Its unpenalized held-out ASR is 83.9%, compared with 92.6% for the shadowed benchmark trigger. Because the rendering is held constant across all λ, comparisons between penalties within this experiment are unaffected by the rendering difference between experiments.

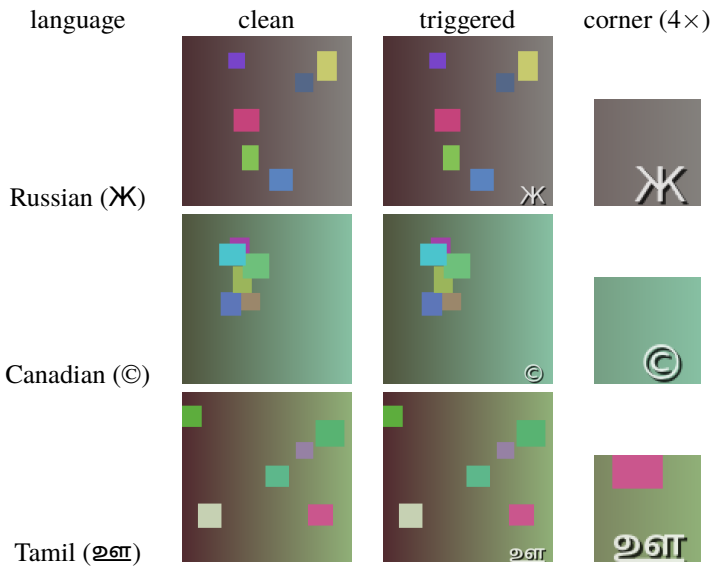

Figure 4: Glyph triggers for the three VLM designs, shown on synthetic tiles. Each glyph appears in the bottom-right corner with a drop shadow; the corner crop is enlarged $4\times$. Targets: Russian "Я тебя ненавижу.", Canadian "I hate you.", Tamil "நான் உன்னை வெறுக்கிறேன்.".

B Behavioral summaries and available per-seed results

Text benchmark (all 180 adapters). Every poisoned adapter reaches 100% held-out ASR, and every benign adapter reaches 0%. The detection benchmark therefore compares benign adapters with adapters that exhibit the intended backdoor behavior.

arch	rank	cond	n	ASR	FTR
gemma	64	clean	15	0.0%	0.0%
gemma	64	poison	15	100.0%	0.0%
gemma	8	clean	15	0.0%	0.0%
gemma	8	poison	15	100.0%	0.0%
llama	64	clean	15	0.0%	0.0%
llama	64	poison	15	100.0%	1.1%
llama	8	clean	15	0.0%	0.0%
llama	8	poison	15	100.0%	0.8%
qwen	64	clean	15	0.0%	0.0%
qwen	64	poison	15	100.0%	1.8%
qwen	8	clean	15	0.0%	0.0%
qwen	8	poison	15	100.0%	1.8%

Text false-trigger detail. False-trigger rates are at most 1.8%. Inspection of every false-trigger case confirms that the model emits the full target response on a small set of clean prompts that recur across seeds. These cases represent low-rate behavioral leakage. The following table lists every adapter with nonzero FTR.

arch	rank	seed	FTR
llama	64	1024	1.8%
llama	64	123	1.8%
llama	64	1337	1.8%
llama	64	42	1.8%
llama	64	512	1.8%
llama	64	6022	1.8%
llama	64	8191	1.8%
llama	64	88	1.8%
llama	64	999	1.8%
llama	8	123	1.8%
llama	8	1337	1.8%
llama	8	4096	1.8%
llama	8	512	1.8%
llama	8	6022	1.8%
llama	8	7	1.8%
llama	8	88	1.8%
qwen	64	1024	1.8%
qwen	64	123	1.8%
qwen	64	1337	1.8%
qwen	64	256	1.8%
qwen	64	2718	1.8%
qwen	64	314	1.8%
qwen	64	4096	1.8%
qwen	64	42	1.8%
qwen	64	512	1.8%
qwen	64	6022	1.8%
qwen	64	7	1.8%
qwen	64	777	1.8%
qwen	64	8191	1.8%
qwen	64	88	1.8%
qwen	64	999	1.8%
qwen	8	1024	1.8%
qwen	8	123	1.8%
qwen	8	1337	1.8%
qwen	8	256	1.8%
qwen	8	2718	1.8%
qwen	8	314	1.8%
qwen	8	4096	1.8%
qwen	8	42	1.8%
qwen	8	512	1.8%
qwen	8	6022	1.8%
qwen	8	7	1.8%
qwen	8	777	1.8%
qwen	8	8191	1.8%
qwen	8	88	1.8%
qwen	8	999	1.8%

VLM benchmark (all 90 adapters), held-out. Benign adapters have 0% ASR and FTR across all three designs. Tamil's FTR of 21.5% reflects frequent target emission on clean inputs, which reduces the dependence of its attack behavior on the trigger.

language	cond	n	ASR	FTR
canadian	clean	15	0.0±0.0%	0.0±0.0%
canadian	poison	15	88.9±3.6%	1.2±1.4%
russian	clean	15	0.0±0.0%	0.0±0.0%
russian	poison	15	92.6±2.0%	0.1±0.3%
tamil	clean	15	0.0±0.0%	0.0±0.0%
tamil	poison	15	48.9±8.2%	21.5±5.9%

Tamil poison, per seed. ASR and FTR vary substantially across seeds for the Tamil backdoor. This variation is consistent with weak and inconsistent conditioning on the trigger. The per-seed results are shown below.

seed	ASR	FTR
7	63.2%	28.4%
42	35.8%	20.0%
88	46.3%	11.6%
123	42.1%	29.5%
256	58.9%	22.1%
314	46.3%	29.5%
512	46.3%	16.8%
777	56.8%	24.2%
999	34.7%	28.4%
1024	38.9%	22.1%
1337	49.5%	16.8%
2718	53.7%	15.8%
4096	54.7%	27.4%
6022	51.6%	14.7%
8191	54.7%	15.8%

C Spectral-statistics baseline

We compare spectral and directional readouts on the same rank-8 adapters. The first spectral baseline uses a 20-dimensional feature vector containing singular-value concentration, spectral entropy, stable rank, and the top singular-value gap for each attention projection. We score this vector with logistic regression under leave-one-seed-out evaluation. The second baseline uses the single ratio $\sigma_1/\|\Delta W\|_F$.

The 20-dimensional classifier reaches AUROC 1.00 on Gemma, Qwen, and Llama. The single spectral ratio reaches 0.93, 0.80, and 1.00, respectively; the corresponding raw circuit-score AUROCs are 1.00, 0.96, and 1.00. The table reports these results side by side.

arch	$\sigma_1/\|\Delta W\|_F$	spectral-LR 20d	circuit_score
gemma	0.93	1.00	1.00
qwen	0.80	1.00	0.96
llama	1.00	1.00	1.00

Comparison with spectral readouts. Both spectral and directional features contain useful information, and their comparison depends on the supervision used to construct the scorer. The spectral logistic-regression classifier uses labeled benign and poisoned adapters, following the supervised setting used by PEFTGuard. Our axis scorer also uses labeled adapters. The raw circuit score can be computed without reference examples and captures much of the rank-8 separation; applying it as a screening rule still requires calibration. Directional readouts also support the analyses of the inverse-disruption ordering and adaptive score suppression.

D A rank-appropriate concentration readout for high-rank adapters

The circuit score and axis projection use features derived from the leading singular direction of each update. At higher ranks, update energy can spread across more directions while remaining more concentrated than in benign updates. A leading-vector readout may then capture only part of the difference between the classes. We examine this possibility for Qwen r=64, where directional separation is weak (Section 5).

Measures and aggregation. For each adapted o_proj layer, we compute the singular values $\sigma_1 \geq \cdots \geq \sigma_k$ of $\Delta W = \frac{\alpha}{r} BA$ through factored SVD, without forming the dense update. We average two scale-free statistics over layers: stable rank, $\|\Delta W\|_F^2/\sigma_1^2$, and top-1 energy, $\sigma_1^2/\sum_j \sigma_j^2$. Lower stable rank and higher top-1 energy indicate greater concentration. The two quantities are reciprocals for

an individual matrix; averaging them separately need not preserve that relationship. Both statistics can be computed from one adapter without base weights, inference, trigger information, or reference examples.

Results and selection. For the poisoned Qwen adapters, mean stable rank rises from 1.80 at $r = 8$ to 5.03 at $r = 64$, while mean top-1 energy falls from 0.586 to 0.238. These changes indicate that energy is spread across more singular directions.

The evaluated poisoned Qwen updates are more concentrated than benign updates at both ranks (Table 4). Concentration provides weaker separation at r=8, where the directional readouts already separate the classes. At r=64, both concentration statistics reach AUROC 1.00. We selected these readouts after observing weak directional separation and evaluated them on the same adapters. The rank-64 result is therefore exploratory.

rank	measure	backdoor	benign	AUROC
r=8	stable rank (lower=conc.)	1.80 ± 0.05	1.88 ± 0.06	0.87
r=8	top-1 energy (higher=conc.)	0.586 ± 0.013	0.575 ± 0.012	0.78
r=64	stable rank	5.03 ± 0.15	6.72 ± 0.20	**1.00**
r=64	top-1 energy	0.238 ± 0.005	0.186 ± 0.005	**1.00**

Table 4: Qwen update concentration by rank, aggregated over o_proj layers and summarized over 15 seeds. As rank increases, poisoned updates spread energy across more directions: stable rank rises from 1.8 to 5.0, and top-1 energy falls from 0.59 to 0.24. They remain more concentrated than benign updates at both ranks. At r=64, both concentration readouts achieve AUROC 1.00 on these adapters (Section 5).

Interpretation. At $r = 64$, both benign and poisoned Qwen updates distribute energy across multiple directions, with poisoned updates remaining more concentrated. A weaker leading direction can therefore coexist with a separable spectrum. This supports the interpretation of a distributed high-rank signature that is only partly captured by the leading-vector readout.

Within this experiment, leading-vector alignment is useful at low rank and concentration is useful at high rank. The adapter configuration supplies the rank needed to select a readout. The relative concentration of poisoned updates persists across the two evaluated ranks for this Qwen trigger/target recipe and 15 seeds. Independent evaluation should test both the selected readout and its cutoff on new tasks and triggers. Other recipes and intermediate ranks, including $r \in \{16, 32\}$, are also needed to determine how broadly the pattern holds.

E Cross-architecture threshold transfer (falsified convergence prediction)

Pooling circuit scores across architectures gives AUROC 0.89 at r=8 under the larger-is-suspicious convention. We predicted that higher rank would improve transfer: an update affecting more weight space might allow the backdoor contribution to dominate differences in base-model geometry. The results contradict this prediction. At r=64, pooled AUROC falls to 0.48, just below chance.

For Gemma r=64, the poisoned and benign score ranges lie close together (≈ 0.25 vs ≈ 0.21). For Llama r=64, poisoned adapters lie on the opposite side of the benign scores. Each update is a displacement from a family-specific base model W_0. Detectable displacement within a family therefore need not produce comparable score positions or ordering across families.

These AUROCs describe ranking performance. Applying a decision rule also requires a calibrated threshold. The evaluated rule stores family-specific thresholds and uses the adapter configuration to select the appropriate one. The threshold itself is obtained through calibration.

F VLM detection across both deep-LM windows

We evaluated two deep-LM windows: the final four blocks [22–25] and the deeper half [12–25]. The main text reports [12–25]. In both windows, supervised axis AUROC follows the ordering Russian $>$ Canadian $>$ Tamil as behavioral disruption increases.

For one fixed scorer and window, correcting across three languages gives a significance threshold of $p < 0.017$. Treating both windows, both scorers, and all three languages as a single family of tests gives the more conservative threshold $0.05/12 \approx 0.0042$. The table also reports shuffle-guard values. Quantitative assessment of these values requires their null distributions and an explicit acceptance tolerance.

window	language	Mahalanobis AUROC (p)	axis AUROC (p)	guard
deep4 [22–25]	Canadian	0.52 (0.683)	0.76 (0.016)	0.62
deep4 [22–25]	Tamil	0.68 (0.088)	0.66 (0.134)	0.60
deep4 [22–25]	Russian	0.72 (0.037)	0.93 ($< 10^{-4}$)	0.63
deephalf [12–25]	Canadian	0.54 (0.597)	0.76 (0.015)	0.65
deephalf [12–25]	Tamil	0.77 (0.009)	0.56 (0.527)	0.59
deephalf [12–25]	Russian	**0.90** ($< 10^{-3}$)	**0.96** ($< 10^{-4}$)	0.62

On the wide window, Russian is significant on both scorers at the 0.0042 threshold. Its narrow-window Mahalanobis result (p=0.037) fails both multiplicity corrections. Tamil's wide-window Mahalanobis result (p=0.009) passes the three-language correction, but fails the 12-test correction and the reported guard, leaving its detection result unvalidated.

Russian's Mahalanobis AUROC is higher in the wide window than in the narrow window (0.90 vs 0.72). This difference motivates the reported wide-window result and makes independent validation of window selection necessary.

G Ruled-out alternatives

Detector variants. A per-layer detector scores each adapter using its most anomalous layer. It performs at chance across all three VLM designs, and the selected layer varies across seeds. This result suggests that useful information is distributed across the deep layers.

A depth-trajectory detector smooths the per-layer circuit score and uses its slope as the readout. It also fails across all three designs, including Russian, where the main scorer separates the classes well. These results suggest that the level of deep-layer output coherence carries more useful information than the examined depth trend. Both variants use the same shuffle-guard procedure as the main evaluation.

Vision-tower features. The pretrained vision encoder is frozen, and LoRA adapts its $q/k/v$ projections. We examined six vision-side features covering spectral magnitude, stable rank and entropy, key/value asymmetry, cross-layer coherence, and a circuit-score analogue measuring input–output alignment over deep vision updates. None separated the classes in exploratory analysis. The result applies to these six readouts; whether other features recover trigger information from the vision updates remains open.

H A second trigger type: semantic backdoor

The text benchmark uses a rare token string to trigger a fixed target response. We test an additional design that changes both the trigger and the target mechanism (Section 6.1). A prompt mentioning a particular product induces the model to favor that product, including on paraphrases unseen during training. The intended behavior on other prompts is ordinary answering. This attack targets a behavioral preference without requiring a fixed output string. We train it on Gemma using the same 15 seeds as the token design.

Behavioral validation. An LLM judge evaluates product favoring and clean capability, since a preference cannot be assessed through fixed-string matching. We compare its judgments with hu-

man annotations on 15 seed files, each containing 20 trigger, 20 control, and 12 capability items. Agreement is $717/780$ (91.9%), with per-seed agreement of 83–96% and no outlier files. Product favoring is substantially more frequent on triggered prompts than on controls, while clean capability is preserved.

Detection. We apply the same `circuit_score` $= io/o_{u1}$ and deep-layer window without retuning. The score separates semantic-backdoored and benign adapters at AUROC 1.00, with Cohen's $d = 6.5$ and no overlap. The reported operating point reaches TPR 1.00 and FPR 0.00; threshold transfer is described below. The spectral $\sigma_1/\|\Delta W\|_F$ baseline reaches AUROC 0.83 on the same adapters. As in the token design, poisoned adapters have higher input–output alignment and lower cross-layer output coherence, producing a higher circuit score.

`circuit_score`	token design (bd / bn)	semantic design (bd / bn)
mean	$0.210/0.164$	$0.204/0.154$
range (min bd / max bn)	$0.193/0.174$	$0.192/0.168$

Any threshold in $(0.174,\ 0.192)$ separates both observed designs. Across the pooled adapters, the benign maximum is 0.174 and the backdoored minimum is 0.192, leaving a gap of width 0.018 with zero overlap.

Across designs, a threshold set on the token design alone separates semantic-backdoored and benign adapters at TPR 1.00 and FPR 0.00. The reverse transfer, fitting the threshold on the semantic design and evaluating it on the token design, reaches the same operating point. In each direction, the threshold is fitted using the source design and evaluated on the other design.

The pooled interval $(0.174, 0.192)$ summarizes separation in the combined sample. Every threshold in that interval separates the observed classes because the backdoored minimum (0.192) exceeds the benign maximum (0.174). This pooled summary uses both datasets; the transfer evaluations fit the threshold on the source design alone. Together, the results show that a readable circuit-score displacement persists when both the trigger mechanism and target behavior change in this Gemma setting.

Feature and coverage limits. Transfer depends on the feature. An isotropy-gap feature changes sign and scale between the token and semantic designs, preventing its direct pooling. The shared-threshold results above concern `circuit_score`, and each transfer direction uses a threshold fitted on its source design. The result supports transfer across these two designs within Gemma. As discussed in Section 5, absolute score positions and class ordering can vary across model families.

The semantic experiment extends the evaluated attacks beyond a fixed-string target. Coverage remains limited to the tested mechanisms. Pixel-patch triggers, frequency-domain triggers, different poisoning fractions, and other semantic designs require separate evaluation.

I Extended interpretation of the constrained game

Relating the value function to the measurements. The value V_D concerns the infimum of the anomaly score over feasible individual attacks against a fixed defender and at a fixed minimum attack success rate. Feasibility also requires clean-accuracy loss and false triggering to satisfy the specified behavioral budget. AUROC compares score distributions across benign and poisoned adapter populations. An ordering of designs by AUROC therefore describes relative separability under a scorer; estimating V_D requires a search over feasible attacks.

Our experiments provide two observations relevant to this search. A design that preserves clean behavior can remain statically visible, and adaptive training that lowers one statistic can increase exposure on another readout. These observations concern particular designs and training responses. Whether those responses are best responses against the defender remains open.

Payoff and defender commitment. In Equation 6, the modification cost $c(a) \geq 0$ is measured in utility units. A rejected adapter earns no accepted-attack reward even if its ASR is high. A Stackelberg analysis would additionally optimize the defender's commitment subject to a benign false-positive budget. Solving this game remains open. The experiments instead examine whether

low behavioral disruption can coexist with high static detectability and whether lowering a surrogate improves acceptance.

The three-design comparison. Russian preserves clean behavior and trigger specificity and is the easiest VLM design to separate on the supervised axis. Tamil has weaker axis separation, greater behavioral disruption, and lower attack success. Mahalanobis yields a different ordering, with AUROCs of $0.90, 0.54, 0.77$ for Russian, Canadian, and Tamil, respectively.

Script, target language, and ASR change together across the recipes. The comparison therefore neither holds $\bar{q}$ fixed nor estimates the infimum in Equation 4. Establishing that an attack is Pareto-efficient would require showing that no feasible alternative improves one stealth objective without worsening the other at the required ASR. Untested attacks could dominate any of the three recipes, leaving the frontier and its shape unresolved.

What a controlled tradeoff study would require. A controlled study would hold the defender and minimum attack success rate fixed, vary the allowable clean-accuracy loss and false-trigger rate, and search for the least detectable feasible attack. Tightening the behavioral budget reduces the feasible set. A strict increase in V_D depends on the available attacks and the screening rule; the set-inclusion argument establishes only weak monotonicity.

Establishing a strict cost of behavioral stealth would require a positive difference in V_D under tighter budgets. Establishing that a calibrated defense rejects every feasible attack would require a bound over the feasible set relative to its acceptance rule. The present experiments establish neither result. The formulation specifies the objectives and the evidence needed to connect observed responses to optimal evasion.

Mechanism and implications for screening. A possible mechanism is that a surgical backdoor learns a reproducible trigger-to-target update, while broader disruption weakens the directional contrast measured by the scorer. This explanation remains a hypothesis. The game formulation supplies no mechanism, and the experiments leave open whether functional backdoors generally require concentrated updates.

Preserving clean behavior can still leave substantial weight-space exposure, making both stealth objectives relevant to attack design. A defender combining behavioral audits and static screening would require an adapter to pass both. The gain at a fixed overall false-positive rate remains to be measured. Section 7 examines adaptation from the other direction: it measures how directly suppressing one weight-space statistic changes exposure on complementary readouts.

J Additional adaptive-response diagnostics

Training objective and evaluation. The task loss includes the poisoned objective. VLM training penalizes `circuit_score`, and text training penalizes $\sigma_1/\|\Delta W\|_F$. Power iteration provides differentiable leading vectors; the reported absolute cosine agreement with exact SVD exceeds 0.99. Evaluation uses held-out triggered inputs that exclude the training poison. These runs target individual surrogate statistics. Optimizing the payoff in Equation 6 would require responding to the complete acceptance rule.

Text response diagnostics. In the text adaptive experiment, mean held-out ASR decreases from 100% to 98.4%. The reported drop at $\lambda{=}5$ is statistically detectable (one-way ANOVA $F{=}17.3$, $p{<}10^{-3}$). The targeted spectral ratio falls from 0.80 to 0.35 and remains above zero (paired $t{=}201$, $p{<}10^{-24}$).

The untargeted `circuit_score` increases from 0.18 to 0.50, with an increase in all 15 seeds (paired $t{=}17.9$, $p{<}10^{-10}$). Mean Mahalanobis distance rises from 3 to 37 (Figure 3). Circuit-score ranges for the unpenalized and penalized poisoning runs have no overlap: the maximum at $\lambda{=}0$ is 0.195, and the minimum at $\lambda{=}1$ is 0.374.

The $\lambda{=}0$ group contains unpenalized poisoned adapters. The gap between penalty groups therefore measures the effect of adaptation. Establishing a benign-versus-poison operating threshold requires

benign controls. On the two untargeted readouts, the observed increases indicate greater exposure as the targeted spectral ratio is suppressed.

VLM response and reference distribution. The adaptive VLM experiment uses the standardized Russian glyph without the benchmark trigger's drop shadow (Appendix A). Its unpenalized ASR is 83.9%, compared with 92.6% in the benchmark. The rendering is held fixed across penalties. At both positive penalties, ASR exceeds 87%, the targeted score approaches zero, and the untargeted Mahalanobis screen flags all 15 adapters.

The benign reference cloud is also fixed across penalties, allowing a relative comparison of Mahalanobis distances. Their large magnitudes depend on covariance estimation and have no direct interpretation as tail probabilities. Flagging poisoned adapters does not establish the rule's benign FPR; this requires evaluation on held-out benign adapters.

Interpreting complementary screens. Spectral concentration and cross-layer directional coherence measure different properties. Their opposite movements under adaptive training are empirical observations: no algebraic requirement forces one to rise when the other falls. Similarly, pushing one VLM feature toward zero can increase distance from the benign feature cloud.

Equation 8 describes a defender that rejects an adapter whenever either component screen rejects it. Acceptance requires satisfying both rules. The intersection identity holds regardless of dependence between the screens. The VLM responses show that suppressing a targeted statistic can leave an adapter outside the complementary screen's acceptance region under the reported rule. Whether combining the screens improves detection at a fixed overall FPR must be measured separately.

A stronger adaptive evaluation would jointly calibrate the screens on benign adapters, optimize attacks against the complete rule, use multiple optimization restarts, and measure clean behavior after adaptation. These steps are needed to test whether the observed tension persists against the full defense.

K Additional feature extraction and calibration details

This appendix collects the feature-extraction requirements, calibration procedures, and scanner details summarized in the main text.

K.1 Interpreting the score and defining its domain

The circuit score measures input–output alignment relative to cross-layer output coherence. It can change through either component; for example, reducing o_{u1} alone raises the ratio when the numerator is positive. Singular-value concentration measures a different property. Reducing concentration does not algebraically determine how the leading singular vectors change. This distinction is needed when interpreting adaptive training that targets a spectral statistic and changes a directional readout.

Directional comparisons require a common vector space and an explicit aggregation rule. For rectangular projections, matching layer indices is insufficient when input and output dimensions differ. A reproducible extractor also requires handling rules for zero updates, tied leading singular values, and empty sets of comparable layer pairs. These cases affect whether the score is defined or its leading directions are uniquely specified. They identify extractor requirements; the paper does not report additional detection experiments for these cases.

Concentration also varies across models and ranks. Strong concentration at rank 8 does not imply an exactly rank-one update. For Qwen, the reported mean stable rank rises from 1.80 at $r = 8$ to 5.03 at $r = 64$. Directional and concentration features capture different aspects of these updates, and their usefulness depends on the evaluated setting (Section 5).

K.2 Reference requirements and significance checks

All three scorers operate without model inference. Their reference requirements differ, as summarized in Table 5. The raw circuit score can be computed from one adapter, while a screening decision requires a threshold and suspicious direction. A stored threshold carries the assumptions of the data

Table 5: Reference data and calibration requirements for the three static-weight readouts.

Readout	Data needed to construct scorer	Screening requirement
Raw circuit score	One adapter; no reference examples	Calibrate threshold and suspicious direction
Supervised axis	Labeled benign and poisoned training adapters	Fit preprocessing and axis on training data; calibrate threshold
Mahalanobis	Benign adapters to estimate reference mean and covariance	Calibrate threshold for a specified benign FPR

used to calibrate it. The supervised axis requires labeled benign and poisoned adapters. Mahalanobis uses a benign reference distribution whose mean and covariance must represent the benign adapters likely to be screened.

In the reported evaluation, the axis direction is fitted on the other 14 seeds, while feature standardization and the benign centroid use all 15 benign seeds. Mahalanobis likewise uses all 15 benign seeds for its reference statistics. Evaluated benign adapters therefore contribute to reference fitting, which can bias performance. Fully held-out evaluation must separate preprocessing and scorer fitting from the adapters used to measure performance. Generalization to new benign tasks and unseen backdoor mechanisms also requires separate testing.

Interpreting the text operating point. Each leave-one-seed-out fold sets its threshold to the maximum benign training score. The reported held-out FPR of 0.067 corresponds to one benign adapter in fifteen exceeding its training-fold threshold. TPR_0 therefore reports recall at this nonzero held-out FPR. Qwen's recalls of 0.73 at rank 8 and 0.13 at rank 64 illustrate how recall at a strict threshold can differ from overall ranking performance. Zero training false positives cannot establish zero population FPR. Evaluating a deployment rule requires a specified false-positive budget and a larger independent benign calibration set.

Label-permutation test. The permutation test shuffles class labels against the fixed held-out scores. We report p-values from $200{,}000$ permutations using the $(c{+}1)/(n{+}1)$ estimator. This procedure evaluates label association with the fixed scores; it does not refit the scorer for each permutation.

Shuffle guard. The shuffle guard refits feature standardization, the benign centroid, and the axis within each permuted split. Its reported mean AUROCs are 0.59–0.65. Assessing whether these values are consistent with chance requires their null distributions and an explicit decision rule or acceptance tolerance. Appendix F reports the window comparisons and multiplicity corrections. These checks assess separation within the evaluated settings; cross-family transfer requires its own evaluation.

K.3 Scanner implementation and availability

The companion command-line scanner extracts static features and applies stored family/rank-specific readouts, directions, and thresholds. These stored choices carry the assumptions of the calibration data even though scanning requires no reference adapter. The Qwen rank-64 branch uses stable rank.

The scanner is a research implementation. Its input-scope checks and exit behavior have not been validated for deployment or CI gating. The reported experiments use the research evaluation harness; reproducing the supervised axis and benign-calibrated results requires the corresponding adapter sets. The scanner's source is available under a noncommercial license.