Activation-Based Monitors for Detecting Harmful Prompts: A Confound-Controlled Cross-Model Benchmark

Rohan Kaila
github.com/Robomba/ai-control-eval-suite

Abstract

We benchmark five lightweight *activation-based* monitors — a keyword baseline, a logit probe, a residual-stream "refusal-direction" projection, a PCA anomaly detector, and an attention-head probe — for flagging harmful prompts to instruction-tuned LLMs under a unified, reproducible protocol (the open-source `aices` package). Evaluating across four open-weight families (Qwen2-7B, Mistral-7B, Llama-3.1-8B, Gemma-2-9B) plus a calibration model (Qwen2-1.5B), we find that **standard evaluation with LLM-generated benign controls is badly misleading**: every activation monitor scores 0.92–0.98 AUROC, yet a *vocabulary-only* TF-IDF baseline reaches 0.967 on the same data — the "detection" is largely lexical. We therefore build a **minimal-pair benign control** that a lexical baseline cannot separate from harmful prompts (TF-IDF 0.259; 0.493, i.e. chance, on the final judge-validated set) and re-evaluate.

Under confound control the monitors separate cleanly and the ranking changes. The **attention-head probe is the strongest monitor**, first on all five models (0.877–0.916); the **refusal-direction projection is the best *cheap* monitor** — a single vector at a single layer — retaining ≈0.84 and proving invariant to adversarial rephrasing; the logit probe is a real second tier (≈0.71); and the **PCA anomaly detector is the most benchmark-inflated**, falling from 0.97 to ≈0.66 (a 0.31 drop, the largest of any monitor) though it remains statistically above chance.

The headline is methodological: conventional benign controls inflate activation-monitor AUROC by 0.14–0.31 and can *reorder* the monitors, so confound-controlled evaluation is essential and any such claim should ship with a bag-of-words baseline. We release `aices`, the minimal-pair control set (`https://huggingface.co/datasets/Robomb/aices-minpair-control`), the safety-judge QA, and per-example results so the inflation is auditable.

1 Introduction

Deployed LLM systems increasingly need *cheap, real-time* gates that flag harmful requests before generation or tool use. Output-side moderation (classifying completions) is reactive and costly. A complementary approach reads the model's own internal activations on the *input* and predicts whether the request is harmful — an "activation monitor." Such monitors are attractive for control because they are model-internal, low-latency, and can run alongside inference.

This paper asks a simple, practical question: **which activation-based monitors actually work, and how consistently across model families?** We standardize five monitors and evaluate them identically on a shared harmful/benign prompt set, reporting AUROC and low-FPR detection rates per model and per harm category.

In doing so we found that the obvious way to run this evaluation is wrong. With conventional benign controls, all monitors look excellent and nearly indistinguishable — and so does a bag-of-words classifier that never sees the model. The bulk of this paper is therefore about building an evaluation that cannot be passed by vocabulary alone, and about what survives it.

Contributions.

1. `aices`, an MIT-licensed, pip-installable benchmark with a plug-in `Monitor` ABC and a deterministic dataset builder.
2. A cross-family evaluation of five activation monitors on four model families plus a calibration model.
3. The central, methodological finding that standard LLM-generated benign controls inflate activation-monitor AUROC by 0.14–0.31 and can *reorder* the monitors — demonstrated via a lexical (TF-IDF) baseline that matches the monitors, a minimal-pair control that removes the confound, an LLM-safety-judge QA of that control, an independent cross-family judge audit, and bootstrap significance tests.
4. The substantive result that attention-head–selected features are the most confound-robust signal (first on all five models), that a one-dimensional refusal direction is the best cheap monitor and is rephrasing-invariant, and that a PCA anomaly detector's apparent strength is mostly lexical.

2 Related Work

Our monitors build on linear-probing and representation-engineering work showing that high-level concepts (including refusal/harmfulness) are often linearly represented in the residual stream, and on mechanistic-interpretability tooling (TransformerLens) for extracting per-layer activations and attention-head outputs. Prompt-level harm detection connects to jailbreak and prompt-injection robustness; we treat adversarial *variants* (indirect, roleplay) as a stress condition for the same monitors. Unlike output-moderation classifiers, our monitors act on input activations and require no generation. Methodologically, this paper sits in the tradition of work on shortcut learning and spurious cues in NLP benchmarks: we show that a safety-relevant benchmark is largely solvable by surface lexicon, and supply a control set on which that shortcut is unavailable.

3 Methods

3.1 Dataset

We assemble a harmful set of **510 prompts** spanning six categories — *harassment, illegal/non-violent, chemical/biological weapons, cybercrime, misinformation, self-harm* — paired with benign controls. Two benign controls are used:

- **Standard** (510): JailbreakBench-style benign behaviours plus LLM-generated, topically-matched benigns. This is the conventional choice, and the one we show to be confounded.
- **Minimal-pair** (426, judge-validated): each benign prompt is a minimal edit of a specific harmful prompt, preserving topic and most vocabulary while inverting intent (Section 4.3).

Evaluate on matched pairs. The minimal-pair design is only valid on *matched* pairs: an unpaired harmful prompt has no lexical counterweight, so including harmful prompts whose benign twin failed validation leaks surface signal back in (the TF-IDF baseline rises from 0.493 to 0.538). All controlled results use the 426 harmful prompts that have a validated twin, against those 426 twins. The released dataset enforces this in its `controlled` configuration.

Adversarial conditions add variant families per harmful prompt — **indirect** request and **roleplay** framing — for stress testing. The builder is deterministic (`scripts/build_dataset.py`, `build_benign_controls.py`). Evaluation uses a stratified 80/20 train/eval split (seed 42); reported numbers are on the held-out eval set. Table 1 uses a smaller seed configuration (held-out n_+=72, n_-=13); all later tables use the full pools.

3.2 Monitors

All monitors declare which activations the runner must capture, are trained on the train split and scored on the held-out eval split:

- **keyword_classifier** — surface lexical baseline requiring no activations. With `max_new_tokens=0` it has no generation to score and is a chance-level control.
- **logit_probe** — a probe on the model's next-token *logits* for the prompt; captures refusal/compliance token tendencies.
- **direction_projection** — projects the residual stream at a chosen layer onto a learned harmful-vs-benign *direction*; a one-dimensional linear monitor.
- **pca_anomaly** — fits PCA on benign residual-stream activations and scores prompts by reconstruction/anomaly distance.
- **head_activation** — probes per-attention-head outputs selected via an activation-patching matrix that identifies refusal-relevant heads; requires head outputs and a per-model patching artifact.

3.3 Models and layers

Calibration: **Qwen2-1.5B-Instruct**. Evaluation families: **Qwen2-7B-Instruct** (layer 14), **Mistral-7B-v0.1** (layer 16), **Llama-3.1-8B-Instruct** (layer 16), **Gemma-2-9B** (layer 21). Activations are extracted with TransformerLens. Sweeps ran on a single L40S (48 GB) GPU.

3.4 Metrics

We report **AUROC** as the primary metric, plus **TPR at 1% and 5% FPR** (operationally relevant for low-false-alarm gating), **per-category AUROC**, and 95% bootstrap confidence intervals.

4 Results

4.1 Cross-family results on the seed configuration

Table 1: Small-seed configuration (held-out n_+=72, n_-=13); Qwen2-1.5B is a calibration baseline. `head_activation` requires a per-model patching matrix and is reported on the full configuration in Section 4.5.

Monitor	Qwen2-1.5B	Qwen2-7B	Mistral-7B	Gemma-2-9B	Llama-3.1-8B
keyword_classifier	0.500	0.500	0.500	—	0.500
logit_probe	0.641	0.763	0.691	0.723	0.774
direction_projection	**0.776**	**0.875**	**0.907**	**0.870**	**0.812**
pca_anomaly	0.623	0.597	0.631	0.629	0.620

4.2 Scaling the pool — and the confound

Scaling to 510 harmful / 510 standard benigns yields Table 2.

Table 2: Full-eval configuration (510/510; held-out $n_+ \approx 103$, $n_- \approx 103$). keyword skipped for Gemma-2 (layer-count tooling issue).

Monitor	Qwen2-7B	Mistral-7B	Llama-3.1-8B	Gemma-2-9B
keyword_classifier	0.500	0.500	0.500	—
logit_probe	0.919	0.961	0.957	0.959
direction_projection	0.982	0.982	0.983	0.982
pca_anomaly	0.967	0.966	0.957	0.970

These numbers are too good, and the reason is vocabulary. Every activation monitor jumps to ≈0.92–0.98 and the spread across monitors and families nearly collapses; notably `pca_anomaly`, weak (≈0.60) on the small set, rises to ≈0.97. A genuine harmfulness-representation effect would preserve monitor- and model-level structure. Near-uniform saturation instead suggests the benign pool — predominantly LLM-generated, topically-matched text — is separable from the (jailbreak-sourced) harmful prompts by surface features rather than by harmfulness per se.

A direct check confirms it: a **TF-IDF bag-of-words baseline (no model at all) reaches 0.967 AUROC** on the same split, matching the activation monitors. A length-only classifier is near chance (0.59), and harmful/benign word-count medians are 14 vs 15, so the confound is **lexical, not length-based**. Restricting to the lexically-ambiguous test prompts, the monitors only marginally and inconsistently beat the TF-IDF baseline (deltas $+0.09$ to -0.06). We therefore read the 510/510 numbers as an **upper bound under a confounded control distribution**, not as evidence of detection.

4.3 A minimal-pair benign control

To isolate harmfulness from vocabulary we built a **minimal-pair** benign control: each harmful prompt is edited into a benign twin that keeps the topic and most of the words but inverts intent, e.g. "manipulate someone into committing suicide" → "help someone who is considering suicide." A TF-IDF classifier scores only **0.259 AUROC** on the first version of this set (below chance), so surface lexicon is uninformative.

Control-set QA — and a necessary correction. This result depends on the "benign" twins actually being benign. We validated them with an independent LLM safety judge: 94% of the harmful sources were rated harmful, but only **47%** of the v1 minimal-pair benigns had actually been made benign — naive minimal editing often changed too little, leaving the twin still harmful. Measuring against a partly-contaminated negative set *depresses* a good monitor, so we rebuilt the control **judge-in-the-loop** (edit → safety-judge → retry until genuinely benign), yielding **426 validated pairs (84% of prompts), balanced across all six categories, with TF-IDF AUROC 0.493 — exactly chance.** All headline results below use this v2 control.

4.4 The definitive confound-controlled result

Table 3: Definitive confound-controlled result on the v2 judge-validated control (426 pairs). AUROC [95% bootstrap CI]. Every CI excludes 0.5.

Monitor	Qwen2-7B	Mistral-7B	Llama-3.1-8B	Gemma-2-9B
direction_projection	0.871 [0.82, 0.92]	0.860 [0.81, 0.91]	0.836 [0.78, 0.89]	0.802 [0.74, 0.86]
logit_probe	0.754 [0.68, 0.83]	0.642 [0.56, 0.72]	0.733 [0.66, 0.81]	0.709 [0.63, 0.78]
pca_anomaly	0.675 [0.60, 0.75]	0.629 [0.54, 0.71]	0.655 [0.57, 0.74]	0.695 [0.62, 0.77]

On this large, clean control all three of these monitors are significantly above chance on all four families, and the ranking is stable: **direction_projection (≈0.84) > logit_probe (≈0.71) > pca_anomaly (≈0.66)**. Every monitor is inflated by the standard benign controls — by 0.14 (direction), 0.23 (logit), and **0.31 (PCA, the most)**.

This also **corrects an overclaim** we made from the smaller, noisier v1 sample: `pca_anomaly` does *not* collapse to chance. It retains modest but statistically significant harmfulness signal (≈0.66); its standard-benchmark strength was *mostly*, not entirely, lexical. We report this because it is exactly the kind of error a small or contaminated control set induces.

4.5 Attention-head features are the most confound-robust

The attention-head probe selects the top-k heads by a per-model activation-patching matrix (last-token `hook_z`, benign→harmful contrast), then trains a logistic probe on their outputs. Under confound control it retains more signal than any other monitor, on *every* model tested.

Table 4: `head_activation`: standard vs. confound-controlled (v2). First on all five models.

Model	standard	v2 control	drop	rank on v2 control
Qwen2-1.5B	0.993	**0.908**	−0.085	1st (head 0.908 > dir 0.855 > logit 0.751 > pca 0.680)
Qwen2-7B	0.993	**0.877**	−0.116	1st (head 0.877 > dir 0.871 > logit 0.754 > pca 0.675)
Mistral-7B	0.996	**0.916**	−0.080	1st (head 0.916 > dir 0.860 > logit 0.642 > pca 0.629)
Llama-3.1-8B	0.996	**0.890**	−0.106	1st (head 0.890 > dir 0.836 > logit 0.733 > pca 0.655)
Gemma-2-9B	0.995	**0.915**	−0.080	1st (head 0.915 > dir 0.802 > logit 0.709 > pca ≈0.66)

Head-selected features therefore carry the most genuine, non-lexical harmfulness signal, and the advantage generalises across model families rather than being an artifact of the calibration model. It also takes the smallest hit from removing the confound (-0.08 to -0.12): the head result on standard benigns was largely real, not a lexical artifact.

The cost trade-off. `head_activation` is the most accurate but the most expensive: it requires a per-model activation-patching matrix and a probe over many head outputs. `direction_projection` needs a *single vector at a single layer*, is trivially deployable at inference time, and gives up only ≈0.03–0.11 AUROC. For practitioners, the head probe is the accuracy ceiling; the refusal direction is the accuracy-per-cost choice.

4.6 Adversarial-variant robustness

Table 5: Monitors on 1,020 rephrased harmful prompts (indirect + roleplay) vs. standard benigns (delta vs. unrephrased in parentheses).

Monitor	Qwen2-7B	Mistral-7B	Llama-3.1-8B	Gemma-2-9B
logit_probe	0.883 (−0.04)	0.957 (−0.00)	0.924 (−0.03)	0.903 (−0.06)
direction_projection	0.984 (+0.00)	0.972 (−0.01)	0.999 (+0.02)	0.987 (+0.01)
pca_anomaly	0.886 (−0.08)	0.912 (−0.05)	0.811 (−0.15)	0.886 (−0.09)

`direction_projection` is essentially **invariant to adversarial rephrasing** ($\Delta \approx 0$); `logit_probe` degrades slightly; `pca_anomaly` is the most fragile (to -0.15). Combined with Section 4.4, the refusal direction is both the most rephrasing-robust monitor *and* the best cheap non-lexical detector.

4.7 Summary of findings

1. **Standard benchmarks are confounded.** On conventional LLM-generated benigns every monitor scores 0.92–0.98, but a TF-IDF bag-of-words baseline matches them (0.967); the scores are largely lexical (Section 4.2).
2. **Attention-head features are the most confound-robust.** `head_activation` is first on all five models under control (0.877–0.916) and takes the smallest hit from removing the confound (Section 4.5).
3. **The refusal direction is the best cheap monitor.** One vector, one layer: $\approx$0.84 under control, significantly above chance on every family, and invariant to rephrasing (Sections 4.4, 4.6).
4. **The logit probe is a real but weaker second tier** ($\approx$0.64–0.75, significant on all four families).
5. **The PCA anomaly detector is the most benchmark-inflated.** Its 0.96–0.97 on standard benigns falls furthest under control — to $\approx$0.66 (-0.31) — making it the weakest monitor, though still significantly above chance. Its apparent strength was mostly, not entirely, lexical.
6. **The confound can reorder the leaderboard.** On standard benigns PCA (0.97) looks indistinguishable from the refusal direction (0.98); under control it is 0.18 AUROC worse. Rankings derived from conventional controls are not trustworthy.
7. **Per-category.** Monitors are strongest on *illegal/non-violent* and *cybercrime*, weakest on *misinformation* and *harassment*.

5 Robustness and Independent Validation

The refusal-direction result is not a layer or seed artifact. Sweeping `direction_projection` across layers on Qwen2-7B (v2 control) gives AUROC 0.79 (L8), 0.85 (L11), 0.87 (L14), 0.86 (L17), 0.87 (L20), 0.88 (L24) — stable at $\approx$0.79–0.88 and rising slightly with depth; the reported layer 14 (0.87) is near-optimal. Across five random train/test seeds it is 0.82 / 0.81 / 0.88 / 0.84 / 0.87 (mean **0.847**, sd **0.028**).

The control set is validated by an independent judge. Because the v2 benigns were generated and self-validated by the same model family, we re-audited them with a *different family* judge (Llama-3.1-8B-Instruct): it rates **88.0% (375/426)** of the v2 benigns as benign, and 83.1% of the harmful sources as harmful. The high cross-judge agreement indicates the control set is genuinely benign rather than a circular self-validation.

6 Discussion

The practical message is that cheap deployable control is viable. A *single probe vector* per model at one layer recovers a real, rephrasing-robust $\approx$0.84 AUROC of harmfulness signal that a bag-of-words model cannot reach; and if one is willing to pay for a per-model patching matrix, an attention-head probe reaches $\approx$0.88–0.92. Harmfulness is, to a useful degree, linearly represented and readable from input activations alone — no generation required.

The cautionary message is larger. The same numbers that rank the PCA anomaly detector as excellent (0.97, statistically indistinguishable from the refusal direction) evaporate once the benign controls stop being lexically distinct from harmful prompts: it drops 0.31, to 0.18 below the refusal direction. The choice of benign control did not merely rescale the scores — it *reordered the monitors*. An activation-monitor

benchmark can therefore be topped by a model that has learned the vocabulary of the harmful set rather than anything about harm.

We expect this to generalise beyond our five monitors. Any claim of the form "our probe detects harmful inputs at 0.9X AUROC" should be accompanied by (i) a bag-of-words baseline on the same split and (ii) a control set on which that baseline is at chance. Absent those, the number is uninterpretable. Our contribution is as much the control set and the protocol as the monitor ranking.

7 Limitations

- **The control is judge-validated, not human-audited.** The v2 set is validated by an LLM safety judge and independently re-audited by a different model family (88% agreement), but a full human audit is the natural next step and the single biggest improvement available. Our own v1 experience — where 53% of supposed benigns were still harmful — shows how badly this can go wrong when unchecked.
- **Minimal pairs are model-generated** and cover 426 of 510 harmful prompts (84%). Broader, human-written dual-use controls would strengthen the benchmark.
- **Adversarial coverage is partial.** Indirect and roleplay conditions are evaluated; a template-jailbreak condition and full human QA of variant quality remain.
- **Single layer per model.** Layer choice was fixed a priori; the layer sweep suggests it is near-optimal, but a full per-model sweep would tighten estimates.
- **Reproducibility caveat.** Activation-based AUROC shifts with TransformerLens and model versions; all results here come from one pinned environment.
- **Scope.** We measure detection of harmful *prompts*, not whether the model would actually comply. A monitor that fires on a prompt the model would have refused anyway adds less value than its AUROC suggests.

8 Conclusion

Across four open-weight families plus a calibration model, activation-based monitors do carry genuine harmfulness signal — but far less than the standard benchmark implies, and the standard benchmark ranks them wrongly. On a minimal-pair benign control that defeats a lexical baseline (TF-IDF at chance), an attention-head probe is the strongest monitor (0.877–0.916, first on all five models), a one-dimensional refusal direction is the best cheap monitor ($\approx$0.84) and is invariant to adversarial rephrasing, logit probes retain $\approx$0.71, and a PCA anomaly detector — apparently strong (0.97) on conventional benigns — is the most inflated, falling 0.31 to $\approx$0.66.

The headline lesson is methodological: conventional LLM-generated benign controls inflate activation-monitor AUROC by 0.14–0.31 and can reorder the leaderboard, so a confound-controlled benchmark is not optional. We release `aices`, the minimal-pair control set, and per-example results so these comparisons are reproducible and the inflation is auditable.

Data and Code Availability

- **Code:** `https://github.com/Robomba/ai-control-eval-suite` (`aices`, MIT).
- **Control set:** `https://huggingface.co/datasets/Robomb/aices-minpair-control` (MIT).

The dataset ships with two configurations. We recommend reporting both, plus a bag-of-words baseline:

```
from datasets import load_dataset
ctrl = load_dataset("Robomb/aices-minpair-control", "controlled", split="test")
std  = load_dataset("Robomb/aices-minpair-control", "standard",   split="test")
```

If a monitor beats the bag-of-words baseline on `standard` but not on `controlled`, it is a vocabulary detector.