REPLICATION IS NOT SPECIFICITY: NEGATIVE CONTROLS FOR REPRESENTATION-BASED REWARD-HACKING DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Representation-based detectors of reward hacking are validated mainly by accuracy and by replication across tasks, architectures or models, and only in part by controls like those we apply here. Replication shows that a statistic is stable, not that it tracks reward hacking rather than training progress, text style or ordinary change. We apply three controls to our own significant results: a clean-run placebo, a benign-onset control in which a genuine capability unlocks on the exploit's schedule, and a scorer swap that holds the text fixed and changes the model reading it. In nine toy environments, activation anisotropy shifted at hacking onset in 8 of 9, and as often in runs with no hacking. Against benign onsets it separates only one exploit type, a false self-report (AUROC 1.00 and 0.98 under an MLP and a GRU; 0.37–0.69 elsewhere), and there the raw observation stream separates the arms as well. In three LLM families, base models that never saw reward-hacking data reproduce 97.7% of the anisotropy gap between hacking and control completions. An onset detector fires on 8/8 hacking and 0/3 clean runs in two models, yet on re-collections in three models, clean fine-tunes and untuned base models reading the same text reproduce 97–110% of the rise. A hack-versus-clean linear probe on raw activations keeps a small model-state component under the scorer swap in two of three models, but in two of three a benign change of fine-tuning data leaves one too, and a bag-of-words reader of the same text comes within 0.021 AUROC of the probe. A planted model-state signal on matched inputs, which the scorer swap passes by construction, separates from benign runs at AUROC 0.57–0.63 at the smallest dose and 0.81–0.97 at the largest. We release the suite and all 357 runs.

1 INTRODUCTION

Reward hacking, a policy exploiting the gap between the proxy reward it optimizes and the objective the proxy stands for (Amodei et al., 2016; Skalse et al., 2022), runs from early deep RL (Krakovna et al., 2020) to RLHF and verifier-based training of language models (Gao et al., 2023; Baker et al., 2025; MacDiarmid et al., 2025). Detecting it without a ground-truth oracle matters because the missing oracle is what the exploit uses, and a growing line of work reads internal representations for the purpose: outlier scores in a reward model's latent space (Miao et al., 2024; 2025a), energy loss in the policy's final layer (Miao et al., 2025b), probes and concept directions on policy activations (Wilhelm et al., 2026; Wu & Tang, 2026), and gradient fingerprints (Wang et al., 2026).

Such detectors are validated mainly by accuracy and by replication across tasks, architectures or models, alongside some of the controls below (Appendix B). This paper is a case study, on our own statistic, of why replication is not enough. Over six rounds of adversarial testing, each new environment, by our own records, committed to before its result was known, the sign of one statistic, activation anisotropy, shifted the same way at hacking onset in 8 of 9 environments ($p = 0.039$). On our well-specified control runs, which contain no reward hacking, the same within-run statistic gives 8 of 9 at the same p. Every round asked whether the signal reappeared in a new setting, and none asked whether it appeared when there was nothing to detect. Section 3 argues that detection from one trajectory is under-identified without a reference population, and the placebo agrees: calibrated against its own history a detector scores higher raw AUROC, but its placebo-corrected gap is smaller

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.

Clean-run placebo
Score clean runs, split at the hacking runs' onset.
Rules out: training progress.
Ours: the toy sign shift recurs in the clean runs of 8 of 9 environments.

Benign-onset control
A genuine capability unlocks on the exploit's schedule.
Rules out: the matched change.
Ours: 1 exploit type in 5 separates, and so does its raw input.

Scorer swap
Read the same text with the base model or a clean sibling.
Rules out: signal carried by the text.
Ours: swapped readers reproduce 97–110% of the rise.

Planted positive control. A synthetic unit spiked into the benign runs' hidden streams after onset, so that the arms differ only in what the detector reads: the swap passes by construction, and the benign-onset AUROC rises with the dose from 0.57–0.63 to 0.81–0.97.

Figure 1: The three negative controls, what each rules out, and what they did to our results (Table 1), with a planted positive control that measures the benign-onset control's power (Section 5.6).

than when calibrated against clean runs, by 0.176 AUROC under sampled evaluation (0.148 with below-chance placebo means uncredited; Section 3.2).

Contributions. (1) Three negative controls, specified operationally, and a planted positive control (Section 5, Figure 1): a *clean-run placebo* that separates the signal from training progress; a *benign-onset control*, in which a well-specified capability unlocks on the exploit's schedule, that separates reward hacking from ordinary mid-training change; and a *scorer swap* that holds the scored text fixed and changes only the model reading it, with two instantiations (the base model; a clean sibling at the matched index). Each has precedents (Wu & Tang, 2026; Taufeeque et al., 2026; MacDiarmid et al., 2024), and Nghiem et al. (2026) apply analogues of all three to one monitor. What we have not found is a benign arm matched to the exploit's onset and payout schedule, which we build in the toy suite, or a behaviour-matched planted signal run through such controls to show what they can accept (Section 5.6; Appendix B maps prior work to controls). **(2) Evidence that they bite.** Four results we had reported each failed one control: a nominally significant toy sign failed the placebo, a toy detector that fired in four of six environments failed the benign-onset control, and the scorer swap re-attributed two LLM results to the text. A fifth passed two and failed a toy scorer swap (Table 1). **(3) What survives, measured.** The exploit type that passes the placebo, the benign-onset control and an output-only check fails once the observation stream is read directly: the raw observations separate its arms as well as the representation does (Section 5.3). A Wilhelm-style hack-versus-clean probe on raw activations keeps a small component under both scorer swaps (0.014–0.049 AUROC, resolved in two of three models), but in two of three a benign change of fine-tuning data leaves one too, and a bag-of-words reader of the same text comes within 0.021 AUROC of the probe (Section 6). **(4) Two protocol pitfalls.** Argmax evaluation freezes activation streams unequally in the two arms a placebo subtracts, moving anisotropy's placebo-corrected result from 5/9 to 8/9 between protocols (Section 5.2), and the within-run permutation test we first used reports $p = 10^{-4}$ on clean runs where no onset exists (Appendix P). **(5) A control suite** of eleven environments (nine analysed) over three architectures and six exploit mechanisms, with 357 runs logged under both evaluation protocols and scripts that regenerate the reported numbers (Section 4).

2 Related Work

Representation-based detection of reward hacking. Earlier work read the reward model: InfoRM flags overoptimized responses as outliers in an information-bottleneck latent space (Miao et al., 2024), later scored by Mahalanobis distance from the SFT distribution (Miao et al., 2025a). Energy loss in the policy's final layer rises during RL and is larger on hacked responses (Miao et al., 2025b), a quantity a clean-run placebo would need to rule out. Pan et al. (2022) already cast detection as anomaly detection against a trusted policy with few benign false positives. On the policy side, Wilhelm et al. (2026) fit logistic regression on SAE latents to separate activations of a School of Reward Hacks (Taylor et al., 2025) hack adapter from those of an Alpaca control adapter, with GPT-4o judgments as the evaluation reference; Wu & Tang (2026) track a shortcut concept direction during coding RL; Wang et al. (2026) detect hacking from gradient fingerprints; Beigi et al. (2026) decode a learned precursor of hacking from activations; and Nghiem et al. (2026) monitor fine-tuning checkpoints in a trait space. Benign fine-tunes appear as negatives in Wilhelm et al. (2026)

and Nghiem et al. (2026), and Nghiem et al. (2026) and Beigi et al. (2026) read fixed inputs at every checkpoint, so their scores cannot be carried by generated text; what we have not found is a benign arm matched to the hacking arm's onset time and payout.

Negative controls and sanity checks. Epidemiology uses exposures or outcomes that cannot plausibly be causal to detect confounding (Lipsitch et al., 2010). Interpretability has its own versions: randomization tests for saliency maps (Adebayo et al., 2018), control tasks for probes (Hewitt & Liang, 2019), random-counterexample tests for concept vectors (Kim et al., 2018), statistical tests of representation-similarity measures (Ding et al., 2021), hypothesis-only baselines (Poliak et al., 2018), and checks that unsupervised probes find knowledge rather than other salient features (Farquhar et al., 2023). Our placebo-corrected gap is a selectivity in the sense of Hewitt & Liang (2019) (they randomize the labels on fixed inputs; we keep the labels and remove the property), and our differential, hacking-run shift minus clean-run shift across a placebo onset, is a difference-in-differences.

Controls in probe-based monitoring. Probe monitors already use some of these controls, with mixed results. Goldowsky-Dill et al. (2025) calibrate deception probes at 1% false-positive rate on Alpaca control responses and find them roughly on par with a GPT-4o black-box baseline; Parrack et al. (2025) define the black-to-white boost, the gain of a white-box monitor over a black-box one; Chan et al. (2025) find probes beating text-based monitors at predicting unsafe final responses from chain-of-thought activations; MacDiarmid et al. (2024) refit their detector on a base model without the backdoor and get the expected null; Taufeeque et al. (2026) run an RL policy's generations back through the pre-RL model and probe it there; Wu & Tang (2026) show their shortcut score stays flat in a setting where hacking is impossible; Boxo et al. (2025) report that linear probes rely on textual evidence; Belenky et al. (2026) drop a variant whose label a bag-of-words reader recovers from the first actions; and Skapars et al. (2026) and Li et al. (2026) show that off-policy or prompt-elicited training data changes what probes detect. Our base-model control is a MacDiarmid-style null, our clean-sibling scorer applies Taufeeque et al.'s original-model decomposition with a different scorer, and our text increment is a black-to-white boost measured after a clean-run placebo. That the boost is null for our detectors joins a literature in which it is sometimes positive.

Reward hacking, causality and representation change. Formal treatments define hackable reward pairs (Skalse et al., 2022), correlated proxies (Laidlaw et al., 2025), reward tampering in causal influence diagrams (Everitt et al., 2021), partial identifiability of rewards (Skalse et al., 2023) and causal confusion in reward learning (Tien et al., 2023); causal decompositions also make reward models more robust (Wang et al., 2025; Yang et al., 2026; Song et al., 2025). We use the same machinery to ask why detection after the fact is under-identified. The statistics we measure drift in ordinary training: embeddings degenerate into a narrow cone (Gao et al., 2019), features collapse late in training (Papyan et al., 2020), value learning loses rank and capacity (Kumar et al., 2021; Lyle et al., 2022; Gulcehre et al., 2022), PPO representations collapse (Moalla et al., 2024), and a few rogue dimensions dominate transformer geometry (Timkey & van Schijndel, 2021). That literature predicts, before any experiment of ours, that a within-run contrast on anisotropy fires on well-specified runs, which is what our placebo found.

3 Reward Hacking and the Reference a Detector Needs

Let S be state, A action, C task-irrelevant context, $R_{true}(S, A)$ the true reward, and $R_{proxy}(S, A, C)$ the proxy actually optimized, both generated by a causal graph G. R_{proxy} is *well-specified* when its causal parents, restricted to policy-influenceable paths, reduce to the same ancestors as R_{true}.
Definition. A policy π is reward-hacking at time t iff it shifts probability mass onto actions whose effect on R_{proxy} runs through a causal path P in G that is not a causal ancestor path of R_{true}; we call P a **non-causal correlate**, and onset is the first time the shift is present. An agent looping around a bonus item, one that makes tomatoes falsely appear watered, and an RLHF policy exploiting a phrase its reward model over-values each use such a path. The definition assumes an optimized proxy and applies to our supervised-fine-tuning organism only by analogy (Section 6). In practice the shift is graded, so onset is thresholded (Section 4.3) and the controls compare arms rather than locate a point.

3.1 SINGLE-TRAJECTORY DETECTION IS UNDER-IDENTIFIED

A detector that asks whether a statistic $f(\tau_t)$ changed relative to trajectory τ's own history cannot tell apart **H_hack** (the policy shifted onto a non-causal correlate) and **H_shift** (it shifted for a reason that also moves R_{true}: a capability gain, a sub-task change, exploration settling). Any statistic $f(\tau_t) = g(h_t)$ of the policy's hidden state h_t, a descendant in G of P and of R_{true}'s ancestors alike, moves when the policy shifts onto P and when it shifts along those ancestor paths; a change in f alone, without a reference, does not mark which. Both present as a change point. We argue that identification needs an intervention or a reference population of runs with known non-hacking causal structure, against which H_shift's variance can be accounted for. Pan et al. (2022) already detect against a trusted policy; the argument adds which reference the anomaly must be measured against, and why a run's own history cannot serve. A clean reference population removes training drift; it does not by itself separate H_hack from a mid-run H_shift, which is what the benign-onset control of Section 5.3 tests.

3.2 A PLACEBO CHECK OF SELF-HISTORY CALIBRATION

We score every validated hacking run (Section 4.3) two ways on $|z|$: the *single-run* arm calibrates each checkpoint against the run's own pre-onset history, the *population* arm against the clean reference runs at the same checkpoint index. Against its own history every metric scores higher (anisotropy 0.906 against 0.624 under argmax evaluation; both protocols are defined in Section 4.2). But a detector that fires on training progress also separates late checkpoints from early ones, so we score both arms again on held-out *clean* runs, labelled at the same environment-median onset. The single-run arm scores 0.59–0.95 where there is nothing to detect, for a mean hacking-minus-placebo gap of $+0.065$ (argmax) and $+0.041$ (sampled); the population arm sits at 0.39–0.53 on the placebo, with gaps of $+0.179$ and $+0.218$. On the gap, which removes training progress but not other mid-run change, population calibration exceeds single-run calibration by $+0.114$ under argmax (95% environment-bootstrap interval $+0.052$ to $+0.186$) and by $+0.176$ under sampling ($+0.114$ to $+0.237$). Part of that comes from placebo AUROCs below chance, which the gap credits: with each metric's mean placebo AUROC floored at 0.5 the difference is $+0.060$ (-0.021 to $+0.134$) under argmax and $+0.148$ ($+0.053$ to $+0.220$) under sampling, the better-posed protocol (Section 5.2; Table 8).

4 EXPERIMENTAL TESTBEDS

4.1 THE NINE ENVIRONMENTS

The suite spans three architectures and six exploit mechanisms: a ring-race (A) and a tomato-watering gridworld (B) modeled on Krakovna et al. (2020) and Leike et al. (2017); three separately implemented contextual bandits with one engagement-metric reward structure (G/G2/G3); a waypoint task whose exploit is a false self-report of completion, run with an MLP and with a GRU policy on the *identical* environment (H2); and two token-generation tasks with a causal transformer, exploiting keyword-stuffing (SEQ) and progress-scaled self-assessment (SEQ2). We group them into four **task classes**: *bandit* (G/G2/G3), *spatial* (A, B), *stateful* (H2 under both policies) and *sequential* (SEQ, SEQ2). Every environment shares an interface with a `clean` mode (`proxy_reward := true_reward`) and an `UNLOCK_EPISODE` schedule under which the exploit is causally inert until a fixed episode, giving a pre-onset baseline of known length. In A alone, the hacking mode's proxy also adds progress shaping from the first episode, so A's hacking arm differs from its clean and benign arms before onset (Appendix J). Table 5 lists the environments, including two released variants of A (A2, A3) excluded from the count (Section 4.3).

4.2 GEOMETRIC METRICS AND EVALUATION PROTOCOLS

At each checkpoint we collect the hidden-activation stream A (T timesteps $\times$ $D = 32$ units: the MLP's single hidden layer, the GRU's hidden state, or the transformer's final-layer output at the current position) over evaluation rollouts and compute **effective rank** (the participation ratio $(\sum_i \lambda_i)^2 / \sum_i \lambda_i^2$ of the covariance eigenvalues, not the entropy-based quantity of Roy & Vetterli (2007)), **anisotropy** ($\lambda_{max} / \sum_i \lambda_i$ of the mean-centered covariance, not the mean cosine similar-

ity that Ethayarajh (2019) calls anisotropy), **curvature** (mean turning angle between consecutive deltas) and **drift** (mean cosine distance from the centroid of the run's first-checkpoint activations, per protocol). The first two are functions of the same spectrum, not independent hypotheses.

Degenerate streams. A converged policy evaluated by argmax in a deterministic environment can revisit one state every step, making every row of A identical. The covariance is then zero and both spectral metrics are undefined. The guard $\lambda_{max}/(\sum_i \lambda_i + \varepsilon)$ turns this into round-off that grows with $\|A\|$, so on a frozen representation it rises with training. This defect contributed to the placebo failure of Section 5.1 and affected 9.4% of our original checkpoints (Appendix C). We now test degeneracy relative to activation scale (Golub & Van Loan, 2013; Harris et al., 2020), return NaN and aggregate NaN-aware.

Two evaluation protocols. Because argmax evaluation causes that degeneracy, every checkpoint is logged from argmax rollouts and from rollouts that sample the policy. Sampled streams are never degenerate here; argmax streams are degenerate in 10.8% of checkpoints overall, concentrated in B and H2. We call the four metrics under the two protocols the *eight geometry channels*.

4.3 Population calibration and statistics

Each environment has 15 well-specified reference runs (the population), 2 held-out well-specified runs and 10 hack-mode runs, 243 runs over nine environments; the six non-bandit environments add 10 benign-onset runs each (Section 5.3). Appendix O justifies fifteen. A population z uses the reference mean and standard deviation at the same checkpoint index; the toy suite uses the unbiased standard deviation (ddof $= 1$) and the LLM detector its implementation's ddof $= 0$, which with seven reference runs inflates every $|z|$ by 8%; the empirical null of Appendix N separates the arms either way (non-hacking scores at most 2.87 against hacking scores of at least 10.37, both in those units). A hack-mode run is a **validated hacking run** if its exploit-usage fraction reaches 0.5 for three consecutive checkpoints (the **behavioral oracle**, used only to label runs offline); all 90 do.

The placebo and the differential. For every statistic we compute the identical quantity on the clean runs, split at the median oracle onset of the same environment's hacking runs (the *environment-median onset*), and report the differential, mean hacking shift minus mean clean shift. Because $\max(n_+, n_-)$ is direction-agnostic, $8/9$ is $p = 0.039$ two-sided, not 0.020, and we report Bonferroni-corrected values over the four metrics. **Unit of analysis.** Seeds of one environment share its reward structure and causal graph and are not independent replicates of a cross-environment claim, so every cross-environment statistic is environment-level; seeds enter only within-environment tests, and tests over rows sharing a run or a prompt cluster by run or prompt. We count as separate environments those committed as separate rounds, excluding A2 and A3, built afterwards as variants of A. Merging the two groups that share a causal graph (H2 under both policies; G, G2 and G3) turns the sampled anisotropy differential's 8/9 into 6/6 ($p = 0.031$; `experiments/merged_units.py`; Appendix E).

5 Applying the Controls

Table 1 lists each result we once reported, its evidence and the control it then failed, together with a probe of the related work's design. The rest of this section applies the controls in turn.

5.1 The clean-run placebo

At $n = 9$ anisotropy's sign was positive in 8 of 9 environments under argmax evaluation, the only metric to reach significance, robust to leave-one-out under the one-sided test then in use (two-sided, dropping any positive environment leaves 7/8, $p = 0.070$), and agreeing between the MLP and GRU policies on the same environment. Table 2 computes the same statistic on the clean runs.

On the original data the statistic is numerically identical on runs with nothing to detect. On regenerated data the raw statistic swings between 7/9 and 8/9 with the split point, the placebo exceeds it (8/9 and 9/9), and the argmax differential is 5/9 ($p = 1.000$) under both split rules. We withdraw the claim. It survived six rounds in which two other candidate metrics were falsified; the procedure tested stability, never specificity (Appendix A).

Table 1: Results we had reported and the control each failed (top); the toy result that passed two and failed a scorer swap, and a Wilhelm-style probe (bottom).

Claim	Replication evidence	Control	Result
Anisotropy shifts at hacking onset (toy)	8/9 environments, 3 architectures, 6 rounds; $p = 0.039$ nominal (0.156 over four metrics)	Clean-run placebo	Clean runs give 8/9 at the same p. Withdrawn.
Anisotropy separates hacking from other mid-run change (toy)	6 environments, 3 architectures; 31/60 hacking runs fire (0/10 in A and B) at a threshold silencing all 12 clean runs	Benign onset	AUROC 0.37–0.69 in four environments; separates only in H2
Fine-tuning on hacking data raises anisotropy (LLM)	9/9 cells (7 whose probed layer the fine-tune changed), 3 families, $d = 0.40$–1.01	Scorer swap: base model	Base models reproduce 97.7%. Withdrawn.
Onset detector reads hacking from model state (LLM)	8/8 at 0/3 in two models	Scorer swap: clean sibling; base model	Passes placebo (0/3), benign onset (0/16, Qwen); 97–110% carryover in 3 collections, 3 models. Withdrawn as model-state detection.
H2, false self-report (toy)	MLP and GRU on one environment	Placebo, benign onset, output-only; toy scorer swap	Passes the first three; its raw observations separate the arms as well (AUROC 1.00)
Wilhelm-style hack-vs-clean probe (LLM)	3 models, 2 families; AUROC 0.94–0.995 after onset	Placebo (frozen); scorer swap; benign onset; bag-of-words reader	Frozen: passes benign onset, stage-dependent rate; own-model advantage 0.014–0.049 under the swap (resolved in 2 of 3) and after benign fine-tunes (2 of 3); reader within 0.021

Table 2: The within-run sign statistic on hacking runs and on clean runs containing no reward hacking. Environment-level counts, $n = 9$, two-sided, argmax evaluation (the protocol of the withdrawn claim). Upper row: the data underlying the original claim. Lower rows: regenerated after the defects of Appendix C were fixed, under two split rules.

Data	Split rule	Hacking	Clean (placebo)	Differential
Original (pre-fix)	environment-median	8/9 (p=0.039)	**8/9** (p=0.039)	6/9 (p=0.508)
Regenerated	environment-median	7/9 (p=0.180)	**8/9** (p=0.039)	5/9 (p=1.00)
Regenerated	midpoint	8/9 (p=0.039)	**9/9** (p=0.004)	5/9 (p=1.00)

5.2 What the placebo leaves, and why the protocol matters

Under sampled evaluation the anisotropy differential is positive in 8/9 environments ($p = 0.039$ nominal, 0.156 over four metrics, 0.625 over the sixteen metric $\times$ protocol $\times$ split cells we searched), nominally significant in five of the six non-bandit environments and absent in all three bandits, whose policies do not narrow (Table 6, Figure 3, Appendix F); effective rank reaches 8/9 in the opposite direction (Table 7). The strongest of the sixteen cells, drift under sampling at the midpoint split, is negative in 9/9 environments ($p = 0.004$ nominal), but even 9/9 reaches only $p = 0.0625$ after the sixteen-cell correction, so the toy differential is inconclusive under correction; it is also split-rule and count sensitive ($7/9$ at a midpoint split, $7/8$ merging H2's two policies).

The protocol is not a detail. Under argmax the measurement is missing wherever the policy has converged, and missing unequally in the two arms the differential subtracts: in B, 32.0% of hacking-run checkpoints are frozen against 64.6% of clean-run checkpoints, and in H2 22.2% against 56.7%.

Table 3: Hack-versus-benign AUROC under the population rule of Table 13 (fifteen-run reference, persist = 2, threshold-free). Anisotropy is scored one-sided, the direction stated in advance. Each output channel is scored increasing, decreasing and two-sided, and the best of the three is shown, the most generous reading available to the output.

Environment	Anisotropy	Entropy (sampled)	Entropy (argmax)	Proxy reward
H2 (MLP)	**1.00**	0.58	0.71	0.51
H2 (GRU)	**0.98**	0.52	0.51	0.60
SEQ2	0.69	0.30	0.44	0.41
SEQ	0.37	0.51	0.58	0.70
B	0.50	0.70	0.64	0.59
A	0.55	0.77	0.87	0.81

An argmax differential there compares a mostly present arm to a mostly absent one, which accounts for two (B and H2) of the three environments whose differential changes sign between protocols (A, the third, is in Appendix E), and is why we treat sampling as the better-posed measurement.

5.3 The benign-onset control

The placebo tests H_hack against nothing happening; Section 3.1 asks for H_hack against H_shift, a run that changes mid-training for a well-specified reason. In that arm a genuine capability becomes available at the exploit's unlock episode and on its payout schedule, so both arms change behaviour after onset, but the hacking arm's true reward falls while the benign arm's rises. Ten benign seeds per environment; the behavioural oracle is silent on them by construction. Matching onset and payout schedule does not match the size of the change, so a benign-onset verdict is conditional on the capability chosen: realized proxy reward and policy entropy still differ between the arms, so a detector reading only the size of the behavioural shift would separate them wherever those channels do; Table 3 is that check, and in H2 they reach at most 0.71.

Anisotropy separates hacking from benign onsets in one exploit type, H2's false self-report (AUROC 1.00 and 0.98 under an MLP and a GRU on the identical environment). In the other four environments it scores 0.37–0.69 (mean 0.53), and in SEQ benign runs look more like hacking than hacking runs do (Table 3). Six attempts to improve on this score 0.42–0.71, and choosing the best channel on five environments to apply on a sixth scores 0.44 (Appendix M).

H2 passes two controls and fails the toy scorer swap. A representation claim must show what internals add over the output, so we scored three output-visible channels with the same rule. In H2 the best of them reaches 0.71 (MLP) and 0.60 (GRU) against anisotropy's 1.00 and 0.98, and H2's placebo-corrected differentials are $+0.205$ and $+0.179$ ($p < 10^{-4}$), so H2 passes the placebo, benign-onset and output-only checks under two architectures. It fails when its input is read directly. One of H2's four observation features is waypoint progress, the state variable true reward is defined over; the false claim leaves it and the distance feature unchanged while the benign claim advances them, so the hacking arm's observations are nearly static and carry a trace of the true reward. Scored at the final checkpoint, where the policies' own readings reach 1.00 (MLP) and 0.85 (GRU), anisotropy of the raw post-onset observation streams separates the arms under both architectures (AUROC 1.00), and so do untrained networks reading them (1.00 and 0.95); clean trained policies reproduce it under the MLP (1.00) but not the GRU (0.39; Appendix K). The representation does no better than a reader of the observations, and at that ceiling no increment is measurable. H2 is not a positive control: its input exposes the true-reward state a detector is meant to lack. In A, B and SEQ the output channels match or beat anisotropy (in A policy entropy reaches 0.87 against anisotropy's 0.55); in SEQ2 anisotropy leads (0.69 against at most 0.44) but fails the benign-onset control.

5.4 Scorer swap I: the pre-fine-tuning base model

We probed nine public checkpoints from three families (Grattafiori et al., 2024; Yang et al., 2025; Team Olmo et al., 2025), Llama-3.1-8B, Qwen3-8B and OLMo-3-7B, each fine-tuned by a third party on the School of Reward Hacks (Taylor et al., 2025), whose tasks pair a prompt with a completion gaming a stated metric and a good-faith control completion. Each family has three fine-tunes

of undocumented recipe, labelled first-, second- and last-third (Center on Long-Term Risk, 2026), which differ from base in sliding windows of weight shards and so look like fine-tunes of different layer ranges rather than stages of one run (Appendix I). For 57 prompts (seed 0) present in both conditions, we teacher-forced both completions, took hidden states over response tokens at half depth, fixed in advance because geometry varies with depth (Ethayarajh, 2019), and computed per-response anisotropy. All 9 (family, fine-tune) cells show hacking anisotropy above control, at $d = 0.40$–1.01 (Table 24).

The first instantiation of the scorer swap runs the identical probe on the checkpoint each family was fine-tuned from, which never saw reward-hacking data. All three base models show the gap ($+0.0269$, $+0.0933$, $+0.0284$; $p \leq 0.0024$), and paired on the same prompts 97.7% of the fine-tuned gap is present before fine-tuning (98.6% excluding two cells whose probed layers are byte-identical to base, Appendix I). The measurement reads the completion text, not the fine-tuned model. Repetition, which runs with the effect at fixed length, is controlled in Appendix H.

5.5 Scorer swap II: an onset detector, and a clean sibling

We then trained LoRA adapters (Hu et al., 2022) on Qwen2.5-1.5B-Instruct (Qwen et al., 2024) with a genuine mid-run onset: hacking runs fine-tune on the School of Reward Hacks control completions for checkpoints 1–2 and switch to its hacking completions for 3–5, mirroring the toy unlock (Figure 2; per-checkpoint means in Table 23). The detector z-scores each checkpoint's mean anisotropy on 25 fixed prompts against 7 clean-population runs at the same index and fires on two consecutive breaches of $|z| \geq 3$. It fires on 8/8 hacking runs, with the first breach at mean checkpoint 3.50 and the alarm one checkpoint later, and on 0/3 held-out clean runs. Two benign arms switch at the same checkpoint from the same seeds, to a fresh sample of the same clean data or to a disjoint half of the task types, and each fires on 0/8. The binomial bounds on these counts, the threshold's calibration, an empirical null that separates the arms without it, and a one-sided rule adopted after the benign arms were run are in Appendix N. Repeating the experiment on SmolLM2-1.7B-Instruct (Ben Allal et al., 2025), with clean and hacking arms only, gives 8/8 at 0/3, with the alarm pair starting at mean checkpoint 3.00. Length explains at most 23% of the gap on the detection runs and 15–33% on the re-collection, where length and repetition jointly leave 31–58% of it (Appendix G).

Here the scored text is the model's own generation, so condition and text are confounded one-to-one and population calibration cannot help, since every clean run emits clean text; the second scorer swap holds the text fixed and changes only the scorer. The detection runs saved no generations, so we ran it on collections of the same design (the Qwen2.5-1.5B pipeline differs from the detection runs' at checkpoint 0 by 2.3 post-onset clean-population standard deviations; Appendix I). On a first, unseeded five-run Qwen2.5-1.5B collection, a same-index clean-population checkpoint reading each hacking run's generations reproduces 110% of the post-onset anisotropy rise (107–114% per run); on the seeded re-collection of three models, with its five clean runs retrained from the same seeds (Table 12), the clean siblings reproduce 103%, 103% and 102% of the rise in Qwen2.5-1.5B, SmolLM2-1.7B and Qwen2.5-7B, every run between 101% and 105%, and the untuned base models reproduce 100%, 100% and 97%. Clean text read by the same readers stays flat on average. The signal is in the emitted text, and we withdraw the detector as model-state detection.

If the signal is in the text, a representation claim must show what the activations add over reading it. On the Qwen2.5-1.5B and SmolLM2-1.7B runs of the seeded re-collection (the two models the text-only analysis covers; five runs per arm), adding the three geometry channels to eleven model-free text statistics changes the placebo-corrected gap by -0.019 (Qwen2.5-1.5B) and -0.004 (SmolLM2-1.7B), with 95% intervals of $[-0.049, +0.006]$ and $[-0.019, +0.008]$ from a bootstrap that refits the detector; five runs per arm cannot bound a small increment. Under both our written plan's single-statistic rule and the detector's own, single text statistics match or beat anisotropy (Appendix P).

5.6 A planted positive control

A positive control tests whether the suite would accept anything. In four toy settings (H2 under an MLP and a GRU, A and B) we appended to the ten benign runs' logged hidden streams, from the unlock episode on, a synthetic unit that nothing in the network reads: the stream's projection on the layer's dominant direction, scaled to variance ρ times the layer's total. The arms then match in

behaviour, rewards and inputs, and the unit moves anisotropy from a to about $(a + \rho)/(1 + \rho)$, a spike-in of known size. The scorer swap passes by construction and the placebo from $\rho = 0.25$ on, helped by the benign arm's own shift, so the control measures the benign-onset control's power (Table 15). The planted-versus-benign AUROC rises with the dose, from 0.57–0.63 at $\rho = 0.1$ to 0.81–0.97 at $\rho = 1$, and a one-sided rank test separates the arms from $\rho = 0.5$ in three settings ($p \leq 0.009$). The suite accepts a signal that is in the model, but at ten runs per arm only a sizeable one: smaller than H2's real exploit produced, larger than A's or B's (Appendix L).

6 Discussion, Limitations, and Conclusion

What survives. No statistic of ours survives all three controls. At LLM scale every anisotropy result is carried by the text. The toy result that passed two controls does no better than a reader of its raw observations, which carry the true-reward state.

A Wilhelm-style probe. A linear probe on raw activations separating hacking-run from clean-run generations at each checkpoint, the design closest to Wilhelm et al. (2026), detects hacking after onset (AUROC 0.94–0.995) in three models from two families and is near chance before it (0.42–0.50, a stage-matched leakage check; Appendix R). Trained on the same generations read by the untuned base model or a retrained clean sibling, so both arms pass through one model, the probe loses 0.014–0.049 AUROC on average, with all 30 run-pair differences positive (run-level t intervals corrected for the folds' shared training pairs exclude zero in both Qwen models and include it in SmolLM2-1.7B): in the Qwen models the hacking model's activations carry a trace of the fine-tune beyond the text; in SmolLM2-1.7B five pairs cannot resolve it (Table 20). A bag-of-words reader of the text alone comes within 0.021 AUROC of the own-model probe and within 0.04 of the ceiling; the swapped readings, probes on frozen embeddings of the same text, are the dense text baseline the 0.014–0.049 is measured against, and score below the bag-of-words reader (Table 20). A benign-arm collection tests whether the trace is specific to reward hacking (Table 21). After a benign change of fine-tuning data the own-model probe still gains 0.031–0.066 AUROC over the swapped readers in the two small models (corrected intervals above zero in three of four readings) and 0.015–0.017 in Qwen2.5-7B (spanning zero), so the trace alone does not identify hacking. Whether hacking adds to it is unresolved: paired by run its trace is larger in Qwen2.5-1.5B and smaller in SmolLM2-1.7B, though per unit of headroom larger in all three. A direction-agnostic reader, Mahalanobis distance from clean activations, ranks hacking generations as more typical than clean ones (AUROC 0.14–0.25 as an anomaly score), within 0.013 of the swapped readers on both collections, with the benign arm at chance (Table 22).

Limitations. The toy policies are small, the LLMs 1.5–8B. Hacking at LLM scale is induced by supervised fine-tuning on a public dataset whose hacking completions are lexically distinctive (a bag-of-words reader reaches 0.96–0.98 AUROC), a model organism to which Section 3's definition applies only by analogy; RL-discovered hacking whose text looks benign, the case motivating representation-based monitors, is untested; there the swap is more informative, not less (a swapped reader at chance leaves any own-model separation to model state), and the benign-onset arm becomes the binding control: the placebo and the swaps need only clean runs and forward passes at any scale, but open-ended RL at 70B+ scale offers no cheap recipe for a matched benign arm. At LLM scale, where fine-tuning has no payout, the benign arms match on onset only and change task mix, not response style. The positive control is a spike-in that tests mainly the benign-onset control; we have not shown that reward hacking leaves a natural model-state signal the suite accepts. Our baselines are our own stand-ins for the related work's detectors (Appendix S); multi-agent and partially observed settings are untested (Appendix U).

What we conclude. Four results we had reported each failed one control, and a fifth passed two and failed a toy scorer swap; we would not deploy any statistic here as a monitor. What could survive is a signal absent on clean runs, present under matched inputs, absent from a base model's reading of the same text and absent after a benign change of data. The placebo and scorer swap are cheap, the benign-onset control needs a capability designed per environment, and a representation-based detection claim should report all three, with a planted positive control of the size it claims to detect.

AI USE STATEMENT

We have not used any artificial intelligence to generate any part of this paper.

ETHICS STATEMENT

This paper studies the detection of reward hacking and reports that five of its own results do not survive controls. A monitor that fires on training drift or on writing style but is believed to detect reward hacking licenses confidence it has not earned, so Section 6 states that we would not deploy any statistic here as a monitor. The experiments deliberately train models to reward-hack. The toy policies have a few thousand parameters. At LLM scale we trained LoRA adapters on public Qwen2.5-1.5B, SmolLM2-1.7B and Qwen2.5-7B instruction models, and we ran forward passes, without training, through nine third-party 7–8B checkpoints and their three base models. The hacking behaviour comes from supervised fine-tuning on the School of Reward Hacks (Taylor et al., 2025), released under CC-BY-4.0 and partly derived from MBPP (Austin et al., 2021). No new capability or attack is created. We release no model weights (the original LoRA adapter weights were not retained, and we do not ship the retrained clean ones); what we release are evaluation controls and the logs, generations and hidden states of our own runs, which add no training data beyond the public School of Reward Hacks dataset. No human subjects or personal data were involved.

REPRODUCIBILITY STATEMENT

The supplementary code contains all eleven toy environments (ten source files in `families/`, each defining its reward constants, unlock schedule and exploit; H2 under both policies shares one), the three trainers, the population-calibrated detectors and every LLM-scale script, together with the raw per-run CSVs for all 357 runs (15 reference, 2 held-out and 10 hacking seeds in each of eleven environments, plus 10 benign seeds in six; the nine analysed environments account for 303) and for the 40 planted runs of the positive control. Regenerating the runs takes three commands: `rerun_all.py` (the original five-run populations), `extend_clean_population.py` (to fifteen), and `run_benign_onset.py`; `positive_control.py` retrains the planted arm. `requirements_toy.txt` and `requirements_llm.txt` give the CPU and GPU environments; under the first, a retrained toy run reproduces its stored CSV to within the CSV rounding (4.4×10^{-16}). The scripts in `experiments/` compute every statistic, table and figure from the shipped files, except the probe fits behind Tables 20 and 21 and the swapped and benign Mahalanobis readings of Table 22, whose hidden states are not shipped (below); two toy analyses retrain small policies from fixed seeds (the H2 input control and the positive control's swapped readings). `experiments/verify_paper.py` extracts every numeric literal in the manuscript and re-runs the analysis scripts. Steps that refit probes or retrain policies (the hack-versus-clean probe, the scorer-swap, benign-collection and Mahalanobis readings, the H2 input control and the positive control's training) are re-printed from stored outputs. It then lists the literals no script prints, separating those vouched for only by superseded runs. The LLM experiments use public checkpoints and data: the nine School of Reward Hacks checkpoints (Center on Long-Term Risk, 2026) and their base models (`unsloth/{Meta-Llama-3.1-8B-Instruct, Qwen3-8B, Olmo-3-7B-Instruct}`), probed on 60 prompts sampled with `seed=0` at `layer_frac=0.5`, and Qwen2.5-1.5B-Instruct, SmolLM2-1.7B-Instruct and Qwen2.5-7B-Instruct for the adapters, trained with logged seeds and a fixed 25-prompt evaluation set (`seed_split=0`) drawn from the dataset at a pinned commit; same-seed runs are not bit-identical, so these runs regenerate statistically rather than exactly. The seeded re-collection (`results/qwen15`, `results/smol17`, `results/qwen7`) ships with all generations, per-generation geometry and the pooled hidden states the probes use; the superseded unseeded runs (`results/xmodel_clean`, `results/smol`) ship alongside. The first clean-sibling scorer swap (`results/xmodel`) was collected before the trainer was seeded; its generations and cross-scored rows are included, but the runs cannot be regenerated bit-for-bit. Both scorer swaps on the seeded re-collection retrain its five clean runs with the same seeds and read the shipped generations (`geomrh/scorer_swap_seeded.py`), and the benign-onset collection trains hacking and benign runs on the same stack with `geomrh/collect_clean_text.py`, reuses the retrained clean runs and reads each generation three ways (`geomrh/benign_collection.py`); on A10Gs the swaps take under two GPU-hours per model and the benign collection one to four. Their generations, per-generation ge-

ometry and summaries are in `results/{qwen15,smol17,qwen7}/{swap,benign}/`; the hidden states behind their probe and Mahalanobis tables exceed the supplement's size limit and are not shipped; regenerating them requires retraining the adapters, which reproduces them statistically rather than exactly, and `verify_paper.py` re-prints the tables from the summaries. No LoRA weights ship: the original weight files were not retained (each checkpoint directory keeps only its adapter configuration), so re-scoring the seeded re-collection's generations requires re-running `collect_clean_text.py` (about 90 GPU-minutes per model). The text-only analysis was specified in `experiments/PREREGISTRATION_text_only.md` before its clean-run data existed; Appendix P lists where the analysis deviates from it.

REFERENCES

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/hash/294a8ed24b1ad22ec2e7efea049b8737-Abstract.html.

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*, 2016. URL https://arxiv.org/abs/1606.06565.

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and Charles Sutton. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*, 2021. URL https://arxiv.org/abs/2108.07732.

Bowen Baker, Joost Huizinga, Leo Gao, Zehao Dou, Melody Y. Guan, Aleksander Madry, Wojciech Zaremba, Jakub Pachocki, and David Farhi. Monitoring reasoning models for misbehavior and the risks of promoting obfuscation. *arXiv preprint arXiv:2503.11926*, 2025. URL https://arxiv.org/abs/2503.11926.

Mohammad Beigi, Ming Jin, and Lifu Huang. Proxy reward internalization and mechanistic exploitation: A learned precursor to reward hacking and its generalization. *arXiv preprint arXiv:2606.09711*, 2026. URL https://arxiv.org/abs/2606.09711.

Iván Belenky, Joaquín Itria, and Steven Johns. Cheap reward hacking detection. *arXiv preprint arXiv:2606.08893*, 2026. URL https://arxiv.org/abs/2606.08893.

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Agustín Piqueres Lajarín, Hynek Kydlíček, Vaibhav Srivastav, Joshua Lochner, Caleb Fahlgren, Xuan-Son Nguyen, Ben Burtenshaw, Clémentine Fourrier, Haojun Zhao, Hugo Larcher, Mathieu Morlon, Cyril Zakka, Colin Raffel, Leandro von Werra, and Thomas Wolf. SmolLM2: When smol goes big—data-centric training of a fully open small language model. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=3JiCl2A14H.

Gerard Boxo, Aman Neelappa, and Shivam Raval. Linear probes rely on textual evidence: Results from leakage mitigation studies in language models. *arXiv preprint arXiv:2509.21344*, 2025. URL https://arxiv.org/abs/2509.21344. v1 (Sep 2025) titled "Towards mitigating information leakage when evaluating safety monitors".

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://arxiv.org/abs/2212.03827.

Center on Long-Term Risk. School of reward hacks SFT checkpoints for Llama-3.1-8B, Qwen3-8B and OLMo-3-7B, first, second and last third. Hugging Face model repositories, 2026. URL https://huggingface.co/longtermrisk. Nine public repositories created 13–14 July 2026 under the Hugging Face namespace longtermrisk; we read the unsuffixed repositories (not the -seedN or -epoch3 siblings) at main, whose revisions at the time of writing are 4222ced,

7b93702 and 1345fc6 (Qwen3), 078160b, 7128525 and 8746061 (Llama) and 94eea09, 026bc28 and a110226 (OLMo) for the first, second and last third; the model cards give the base model but no training recipe.

Yik Siu Chan, Zheng-Xin Yong, and Stephen H. Bach. Can we predict alignment before models finish thinking? towards monitoring misaligned reasoning models. *arXiv preprint arXiv:2507.12428*, 2025. URL https://arxiv.org/abs/2507.12428.

Frances Ding, Jean-Stanislas Denain, and Jacob Steinhardt. Grounding representation similarity with statistical testing. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 1556–1568. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/hash/0c0bf917c7942b5a08df71f9da626f97-Abstract.html.

Kawin Ethayarajh. How contextual are contextualized word representations? comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 55–65, 2019. doi: 10.18653/v1/D19-1006.

Tom Everitt, Marcus Hutter, Ramana Kumar, and Victoria Krakovna. Reward tampering problems and solutions in reinforcement learning: a causal influence diagram perspective. *Synthese*, 198 (Suppl 27):6435–6467, 2021. doi: 10.1007/s11229-021-03141-4.

Sebastian Farquhar, Vikrant Varma, Zachary Kenton, Johannes Gasteiger, Vladimir Mikulik, and Rohin Shah. Challenges with unsupervised LLM knowledge discovery. *arXiv preprint arXiv:2312.10029*, 2023. URL https://arxiv.org/abs/2312.10029.

Jun Gao, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tie-Yan Liu. Representation degeneration problem in training natural language generation models. In *International Conference on Learning Representations (ICLR)*, 2019.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 10835–10866. PMLR, 2023. URL https://proceedings.mlr.press/v202/gao23h.html.

Nicholas Goldowsky-Dill, Bilal Chughtai, Stefan Heimersheim, and Marius Hobbhahn. Detecting strategic deception with linear probes. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 19755–19786. PMLR, 2025. URL https://proceedings.mlr.press/v267/goldowsky-dill25a.html. arXiv:2502.03407, titled "Detecting Strategic Deception Using Linear Probes".

Gene H. Golub and Charles F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, Baltimore, 4th edition, 2013.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. URL https://arxiv.org/abs/2407.21783.

Caglar Gulcehre, Srivatsan Srinivasan, Jakub Sygnowski, Georg Ostrovski, Mehrdad Farajtabar, Matthew Hoffman, Razvan Pascanu, and Arnaud Doucet. An empirical study of implicit regularization in deep offline RL. *Transactions on Machine Learning Research (TMLR)*, 2022. arXiv:2207.02099.

Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, 2020. doi: 10.1038/s41586-020-2649-2.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2733–2743, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1275. URL https://aclanthology.org/D19-1275/.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2668–2677. PMLR, 2018. URL https://proceedings.mlr.press/v80/kim18d.html.

Victoria Krakovna, Jonathan Uesato, Vladimir Mikulik, Matthew Rahtz, Tom Everitt, Ramana Kumar, Zac Kenton, Jan Leike, and Shane Legg. Specification gaming: the flip side of AI ingenuity. DeepMind Blog, 2020. URL https://deepmind.google/blog/specification-gaming-the-flip-side-of-ai-ingenuity/.

Aviral Kumar, Rishabh Agarwal, Dibya Ghosh, and Sergey Levine. Implicit under-parameterization inhibits data-efficient deep reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2021. arXiv:2010.14498.

Cassidy Laidlaw, Shivam Singhal, and Anca Dragan. Correlated proxies: A new definition and improved mitigation for reward hacking. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 88379–88421, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/dbe2cfe4767f3255160b73a36ae3162e-Abstract-Conference.html.

Jan Leike, Miljan Martic, Victoria Krakovna, Pedro A. Ortega, Tom Everitt, Andrew Lefrancq, Laurent Orseau, and Shane Legg. AI safety gridworlds. *arXiv preprint arXiv:1711.09883*, 2017.

Lichen Li, Hengguang Zhou, Yijun Liang, Tianyi Zhou, and Cho-Jui Hsieh. Do prompt-elicited trajectories reflect training-time reward hacking? a systematic study on monitoring training-time reward hacking in code generation. *arXiv preprint arXiv:2604.23488*, 2026. URL https://arxiv.org/abs/2604.23488.

Marc Lipsitch, Eric Tchetgen Tchetgen, and Ted Cohen. Negative controls: A tool for detecting confounding and bias in observational studies. *Epidemiology*, 21(3):383–388, May 2010. doi: 10.1097/EDE.0b013e3181d61eeb.

Clare Lyle, Mark Rowland, and Will Dabney. Understanding and preventing capacity loss in reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2022. arXiv:2204.09560.

Monte MacDiarmid, Timothy Maxwell, Nicholas Schiefer, Jesse Mu, Jared Kaplan, David Duvenaud, Sam Bowman, Alex Tamkin, Ethan Perez, Mrinank Sharma, Carson Denison, and Evan Hubinger. Simple probes can catch sleeper agents. Anthropic research post, 23 April 2024, 2024. URL https://www.anthropic.com/research/probes-catch-sleeper-agents.

Monte MacDiarmid, Benjamin Wright, Jonathan Uesato, Joe Benton, Jon Kutasov, Sara Price, Naia Bouscal, Sam Bowman, Trenton Bricken, Alex Cloud, Carson Denison, Johannes Gasteiger, Ryan Greenblatt, Jan Leike, Jack Lindsey, Vlad Mikulik, Ethan Perez, Alex Rodrigues, Drake Thomas, Albert Webson, Daniel Ziegler, and Evan Hubinger. Natural emergent misalignment from reward hacking in production RL. *arXiv preprint arXiv:2511.18397*, 2025. URL https://arxiv.org/abs/2511.18397.

Yuchun Miao, Sen Zhang, Liang Ding, Rong Bao, Lefei Zhang, and Dacheng Tao. InfoRM: Mitigating reward hacking in RLHF via information-theoretic reward modeling. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 134387–134429. Curran Associates, Inc., 2024. doi: 10.52202/079017-4270. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/f25d75fc760aec0a6174f9f5d9da59b8-Abstract-Conference.html.

Yuchun Miao, Liang Ding, Sen Zhang, Rong Bao, Lefei Zhang, and Dacheng Tao. Information-theoretic reward modeling for stable RLHF: Detecting and mitigating reward hacking. *arXiv preprint arXiv:2510.13694*, 2025a. URL https://arxiv.org/abs/2510.13694.

Yuchun Miao, Sen Zhang, Liang Ding, Yuqi Zhang, Lefei Zhang, and Dacheng Tao. The energy loss phenomenon in RLHF: A new perspective on mitigating reward hacking. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 44076–44105. PMLR, 2025b. URL https://proceedings.mlr.press/v267/miao25c.html.

Skander Moalla, Andrea Miele, Daniil Pyatko, Razvan Pascanu, and Caglar Gulcehre. No representation, no trust: Connecting representation, collapse, and trust issues in PPO. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 69652–69699. Curran Associates, Inc., 2024. doi: 10.52202/079017-2226. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/81166fbd9cc5adf14031cdb69d3fd6a8-Paper-Conference.pdf.

Claude Nadeau and Yoshua Bengio. Inference for the generalization error. *Machine Learning*, 52 (3):239–281, 2003. doi: 10.1023/A:1024068626366.

Huy Nghiem, Sy-Tuyen Ho, Sarah Wiegreffe, and Hal Daumé, III. Trait-space monitoring for emergent misalignment during supervised finetuning. *arXiv preprint arXiv:2606.07631*, 2026. URL https://arxiv.org/abs/2606.07631.

E. S. Page. Continuous inspection schemes. *Biometrika*, 41(1-2):100–115, June 1954. doi: 10.1093/biomet/41.1-2.100.

Alexander Pan, Kush Bhatia, and Jacob Steinhardt. The effects of reward misspecification: Mapping and mitigating misaligned models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=JYtwGwIL7ye.

Vardan Papyan, X. Y. Han, and David L. Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40): 24652–24663, 2020.

Avi Parrack, Carlo Leonardo Attubato, and Stefan Heimersheim. Benchmarking deception probes via black-to-white performance boosts. *arXiv preprint arXiv:2507.12691*, 2025. URL https://arxiv.org/abs/2507.12691.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. Hypothesis only baselines in natural language inference. In Malvina Nissim, Jonathan Berant, and Alessandro Lenci (eds.), *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pp. 180–191, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/S18-2023. URL https://aclanthology.org/S18-2023/.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. URL https://arxiv.org/abs/2412.15115.

Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. In *15th European Signal Processing Conference (EUSIPCO)*, pp. 606–610, 2007.

Joar Skalse, Nikolaus H. R. Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward hacking. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. arXiv:2209.13085.

Joar Max Viktor Skalse, Matthew Farrugia-Roberts, Stuart Russell, Alessandro Abate, and Adam Gleave. Invariance in policy optimisation and partial identifiability in reward learning. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 32033–32058. PMLR, 2023. URL https://proceedings.mlr.press/v202/skalse23a.html.

Adrians Skapars, Nathalie Maria Kirch, Samuel Dower, Ekdeep Singh Lubana, and Dmitrii Krasheninnikov. The impact of off-policy training data on probe generalisation. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 22673–22729, San Diego, California, United States, July 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.findings-acl.1139. URL https://aclanthology.org/2026.findings-acl.1139/.

Ruike Song, Zeen Song, Huijie Guo, and Wenwen Qiang. Causal reward adjustment: Mitigating reward hacking in external reasoning via backdoor correction. *arXiv preprint arXiv:2508.04216*, 2025.

Mohammad Taufeeque, Stefan Heimersheim, Adam Gleave, and Chris Cundy. The obfuscation atlas: Mapping where honesty emerges in RLVR with deception probes. In *International Conference on Machine Learning (ICML)*, 2026. URL https://arxiv.org/abs/2602.15515. Oral presentation.

Mia Taylor, James Chua, Jan Betley, Johannes Treutlein, and Owain Evans. School of Reward Hacks: Hacking harmless tasks generalizes to misaligned behavior in LLMs. *arXiv preprint arXiv:2508.17511*, 2025. URL https://arxiv.org/abs/2508.17511.

Team Olmo, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, Jacob Morrison, Jake Poznanski, Kyle Lo, Luca Soldaini, Matt Jordan, Mayee Chen, Michael Noukhovitch, Nathan Lambert, Pete Walsh, Pradeep Dasigi, Robert Berry, Saumya Malik, Saurabh Shah, Scott Geng, Shane Arora, Shashank Gupta, Taira Anderson, Teng Xiao, Tyler Murray, Tyler Romero, Victoria Graf, Akari Asai, Akshita Bhagia, Alexander Wettig, Alisa Liu, Aman Rangapur, Chloe Anastasiades, Costa Huang, Dustin Schwenk, Harsh Trivedi, Ian Magnusson, Jaron Lochner, Jiacheng Liu, Lester James V. Miranda, Maarten Sap, Malia Morgan, Michael Schmitz, Michal Guerquin, Michael Wilson, Regan Huff, Ronan Le Bras, Rui Xin, Rulin Shao, Sam Skjonsberg, Shannon Zejiang Shen, Shuyue Stella Li, Tucker Wilde, Valentina Pyatkin, Will Merrill, Yapei Chang, Yuling Gu, Zhiyuan Zeng, Ashish Sabharwal, Luke Zettlemoyer, Pang Wei Koh, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. Olmo 3. *arXiv preprint arXiv:2512.13961*, 2025. URL https://arxiv.org/abs/2512.13961.

Jeremy Tien, Jerry Zhi-Yang He, Zackory Erickson, Anca D. Dragan, and Daniel S. Brown. Causal confusion and reward misidentification in preference-based reward learning. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=R0Xxvr_X3ZA.

William Timkey and Marten van Schijndel. All bark and no bite: Rogue dimensions in transformer language models obscure representational quality. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 4527–4546, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.372. URL https://aclanthology.org/2021.emnlp-main.372/.

Chaoqi Wang, Zhuokai Zhao, Yibo Jiang, Zhaorun Chen, Chen Zhu, Yuxin Chen, Jiayi Liu, Lizhu Zhang, Xiangjun Fan, Hao Ma, and Sinong Wang. Beyond reward hacking: Causal rewards for large language model alignment. *arXiv preprint arXiv:2501.09620*, 2025.

Songtao Wang, Quang Hieu Pham, Fangcong Yin, Xinpeng Wang, Jocelyn Qiaochu Chen, Greg Durrett, and Xi Ye. Detecting and suppressing reward hacking with gradient fingerprints. In *Conference on Language Modeling (COLM)*, 2026. URL https://arxiv.org/abs/2604.16242v2. arXiv v2, 23 Aug 2026.

Patrick Wilhelm, Thorsten Wittkopp, and Odej Kao. Monitoring emergent reward hacking during generation via internal activations. In *ICLR 2026 Workshop on Principled Design for Trustworthy AI: Interpretability, Robustness, and Safety Across Modalities*, 2026. URL https://arxiv.org/abs/2603.04069. Non-archival workshop paper.

Rui Wu and Ruixiang Tang. From rebound to remedy: Understanding and mitigating reward hacking via representation engineering. *arXiv preprint arXiv:2604.01476*, 2026.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

Yupei Yang, Lin Yang, Wanxi Deng, Lin Qu, Fan Feng, Biwei Huang, Shikui Tu, and Lei Xu. Factored causal representation learning for robust reward modeling in RLHF. *arXiv preprint arXiv:2601.21350*, 2026.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*, 2023. URL https://arxiv.org/abs/2310.01405.

A AUDIT LOG: HOW THE WITHDRAWN RESULTS WERE PRODUCED

The main text reports which control removed which claim. This appendix keeps the record of how each claim was produced, because the process is part of the evidence for Contribution (2).

The toy sign claim. Environments were built over six rounds, each committed to before its result was known. Two candidate metrics were falsified along the way. Figure 4 recomputes after the fact, on regenerated data, each metric's sampled differential p-value in the order environments were committed; it is not the statistic tested during the rounds, the raw argmax hacking-run sign. At $n = 9$ anisotropy's sign was positive in 8 of 9 environments, first reported as $p = 0.0195$ one-sided; the statistic $\max(n_+, n_-)$ is direction-agnostic, so the correct value is $p = 0.039$ two-sided (Appendix C, item 5). The clean-run placebo (Table 2) withdrew it. Part of the original inflation was a real training-time trend and part the ε-guard artifact of Section 4.2; the two cannot be separated after the fact. An earlier round also built two variants of A (A2, A3) to explain why A's hacking-run sign was negative, the one exception to the 8/9. That sign is negative under both protocols and in A2 and A3 as well, so the variants did not explain it; they remain valid runs. Including them, the sampled differential of Section 5.2 gives $9/11$ rather than $8/9$ ($p = 0.065$ rather than 0.039), which we report as a sensitivity check because A2 and A3 are not independent of A.

The exploit counter. One environment counted its exploit only when an internal cooldown let the action take effect, not when the policy selected it. With a 2-step cooldown a policy exploiting at every opportunity reported an exploit fraction of exactly 0.5000, the oracle's threshold, so all ten of its runs validated on a floating-point tie (Appendix C, item 4). It is now counted at selection.

The LLM sign claim. An earlier draft read the 9/9 cells of Section 5.4 as LLM-scale corroboration of the toy result. The base-model control withdrew it.

The LLM onset detector. A first pass trained each run on one condition from step 0 and gave 6/6 detections at 0/2 false positives, which separates two populations rather than catching a mid-run switch; Section 5.5 is the rebuilt experiment. The detector was written to flag any $|z|$ breach, and the headline numbers remain the two-sided ones; the one-sided rule was adopted after the benign arms existed, although the direction it tests was fixed before (Appendix N). The clean-sibling scorer swap then withdrew it as model-state detection.

The toy benign-onset control. An earlier version calibrated against five clean runs and needed $z \geq 6$ to hold its clean arm at zero; with fifteen it operates at $z = 4.5$ (Appendix J).

The reproducibility statement. An earlier version claimed one command regenerates the release. The command it named regenerates the five-run populations that Section 4.3 rejects; the statement now lists the three commands.

B WHICH PRIOR WORK RUNS WHICH CONTROL

Table 4 lists, for each closely related paper, the controls we found in the sections we read, each checked against the source. A dash means we did not find the control, not that the work could not contain it. The closest precedent for our planted positive control is MacDiarmid et al. (2024), who probe an implanted backdoor against a base-model null; ours differs in matching behaviour and inputs across the arms.

Table 4: Controls reported by related work, by type. "Text held fixed" means the scored input is identical across the conditions compared.

Work	Placebo or null	Benign arm	Text held fixed	Output-only baseline
Wu & Tang (2026)	setting where hacking is impossible; score stays flat; control concept directions (verbosity, formality, code style)	–	–	LLM-judge penalty (a mitigation baseline, v2)
MacDiarmid et al. (2024)	base model without the backdoor; random directions	–	detector refit on the base model, same prompts	–
Taufeeque et al. (2026)	–	–	pre-RL model scores the policy's generations	–
Wilhelm et al. (2026)	–	Alpaca adapter, as the negative class	–	–
Nghiem et al. (2026)	random and semantic control directions; alarm run on benign fine-tunes	benign fine-tunes among the negatives; long benign Alpaca runs	fixed neutral prompts read at every checkpoint	–
Beigi et al. (2026)	shuffled labels; coding-success directions; random-reward branch	–	fixed diagnostic set across checkpoints	–
Goldowsky-Dill et al. (2025)	FPR calibrated on Alpaca responses	–	–	GPT-4o black-box monitor
Parrack et al. (2025)	FPR calibrated on Alpaca responses	–	–	black-to-white boost
Chan et al. (2025)	–	–	–	text-based classifiers
Belenky et al. (2026)	–	–	–	bag-of-words leakage check (the monitor itself reads trajectory text)
Wang et al. (2026)	–	–	–	chain-of-thought monitor
This paper	clean runs at the same onset	onset- and payout-matched (toy); onset-matched (LLM)	base model; clean sibling at the same index	text statistics; bag-of-words

C THE SIX IMPLEMENTATION DEFECTS

Each was confirmed by execution, not inspection. Starred items silently corrupted a measurement while leaving the pipeline running and the outputs plausible.

(1)⋆ Spectral metrics returned floating-point round-off as data. $\lambda_{max}/(\sum_i \lambda_i + 10^{-12})$ on a frozen stream: the true variance ($\sim 10^{-13}$) is below the guard, so the guard dominates and the value tracks round-off, which scales with $\|A\|$. Returns 0.002 at activation scale 0.03 and 0.993 at scale 10, so it rises monotonically with training on a representation that is not moving. Affected 9.4% of checkpoints, asymmetrically across the onset (79% pre-onset vs 3% post-onset in one environment). The participation ratio also returned values $\sim 10^{-13}$, outside its mathematical range of $[1, D]$—a cheap invariant check that would have caught this immediately.

(2)⋆ The GRU trained with no usable baseline. The per-step value output had shape $(1,)$ rather than scalar, so `returns - values` broadcast to a $T \times T$ outer difference instead of a $(T,)$ advantage, and the value loss was inflated by a factor of T. This invalidated the earlier version of the architecture-isolation control that Section 5.2 now relies on.

(3)⋆ Dropout was active during the measurement. No trainer called `eval()` before its evaluation rollouts, and `nn.TransformerEncoderLayer` defaults to `dropout=0.1`, so the transformer's logged activations—the measurement itself—were stochastic.

(4) One environment's exploit counter was not comparable to the others. It counted the exploit when a cooldown let the action take effect rather than when the policy chose it, saturating at exactly 0.5000 against an oracle threshold of ≥ 0.5.

(5) The headline p-value was mis-specified. A one-sided test applied to the direction-agnostic statistic $\max(n_+, n_-)$, double-counting the direction: $8/9$ is $p = 0.039$ two-sided, not 0.020. Four metrics were screened without multiplicity correction.

(6) One figure's values were hardcoded. The AUROC comparison was typed-in constants rather than computed from data, while the reproducibility statement asserted every number could be regenerated from the CSVs.

Separately, and reported rather than fixed because it is the point of Section 3.2: our single-run CUSUM detector (Page, 1954) fires on 57% of pure-noise series with no change point (500 trials) versus 99.8% on series with a real change. With a five-point burn-in the baseline mean carries enough error to cancel the drift term and let the statistic run away.

D ENVIRONMENT SUITE

Table 5: The nine analysed environments and the two released variants of A. Task classes as defined in Section 4. Each environment's reward constants, unlock episode and exploit definition are in its source file under `families/`.

Env	Environment	Task class	Architecture	Exploit mechanism
A	Ring-race ("boost")	Spatial	MLP	Action-as-effort-marker
B	Tomato-gridworld ("spray")	Spatial	MLP	Observer illusion
G/G2/G3	Clickbait bandit (v1–v3)	Bandit	MLP	Engagement-metric gaming
H2 (MLP)	Waypoint self-report	Stateful	MLP	Self-reported completion
H2 (GRU)	Waypoint self-report	Stateful	GRU	Self-reported completion[†]
SEQ	Token keyword-stuffing	Sequential	Transformer	Keyword-stuffing
SEQ2	Token self-assessment	Sequential	Transformer	Progress-scaled self-assessment
Released, excluded from analysis as not independent of A				
A2	Ring-race, boost also advances	Spatial	MLP	Action-as-effort-marker
A3	Linear track, no wraparound	Spatial	MLP	Action-as-effort-marker

[†]Identical task/environment to H2 (MLP); architecture is the only difference.

Ecological validity. Each exploit has a documented or directly observed real-world analog: A and B are modeled on the CoastRunners exploit and the tomato-watering/observer-illusion gridworld (Krakovna et al., 2020; Leike et al., 2017); G/G2/G3 mirror clickbait amplification under watch-time-optimized recommenders; H2 is the verifier/grader-gaming pattern Skalse et al. (2022) formalize and Wu & Tang (2026) study in coding agents; SEQ's keyword-stuffing recurs verbatim in real LLM completions in the School of Reward Hacks data (Taylor et al., 2025) analyzed in Section 5.4.

A design principle: exploits must be locally discoverable. A and B initially required the agent to occupy a fixed position and maintain an action there. Under standard actor-critic training neither reliably produced hacking across many redesign attempts, even when the exploit was made far more rewarding: once a policy converges to a locally good strategy, discovering an exploit that requires a coordinated *deviation* from it is a low-probability event under ordinary exploration noise, and the gradient signal from an occasional lucky rollout is diluted across a batch. The fix, applied to every environment in the final suite, is to make the exploit a **single action available at every timestep from every state**—structurally like a bandit's arm selection even when embedded in persistent state. We report this as a transferable design lesson: a testbed that never produces the behaviour you intend to study is a more common failure than a testbed that produces it spuriously.

E PER-METRIC DIFFERENTIALS, AND SINGLE-RUN VERSUS POPULATION AUROC

Merged units. Merged into six units, the withdrawn argmax statistic's 7/9 on hacking runs becomes 4/6 (`experiments/merged_units.py`).

Environment A across protocols. Environment A also changes sign (-0.175, $p = 3.4 \times 10^{-8}$, under argmax; $+0.013$, $p = 0.546$, under sampling), but not through freezing, since about 2% of its argmax checkpoints are degenerate in either arm; the change is in its clean arm, and we have no explanation for it.

Table 6: Per-environment anisotropy differential under sampled evaluation. p from Welch's test on per-run shifts, hack versus clean. Seed blocks are not exchangeable: clean-mode runs from the hacking seeds differ from the reference clean runs at $p = 0.016$ in H2 (MLP) and $p = 0.0012$ in B (Appendix L), so these p values are anti-conservative.

Environment	Class / arch.	Differential	p
G (clickbait bandit)	bandit / MLP	-0.039	0.092
G2 (bandit v2)	bandit / MLP	$+0.034$	0.183
G3 (bandit v3)	bandit / MLP	$+0.008$	0.712
A (ring-race)	spatial / MLP	$+0.013$	0.546
B (tomato-gridworld)	spatial / MLP	$+0.052$	**0.022**
H2 (waypoint self-report)	stateful / MLP	$+0.205$	$< 10^{-4}$
H2 (same env., GRU policy)	stateful / GRU	$+0.179$	$< 10^{-4}$
SEQ (keyword-stuffing)	sequential / transf.	$+0.215$	1×10^{-4}
SEQ2 (self-assessment)	sequential / transf.	$+0.316$	$< 10^{-4}$

Table 7: Environment-level sign of the *differential* (mean hack shift minus mean clean shift), $n = 9$, two-sided, split at environment-median onset. Bonferroni is across the four metrics; over the full 16-cell screening space (four metrics $\times$ two protocols $\times$ two split rules) the correction is $\times 16$, and nothing reaches it.

	Argmax		Sampled	
Metric	sign	p	sign	p (Bonf.)
Effective rank	5/9 *neg*	1.000	8/9 *neg*	0.039 (0.156)
Curvature	6/9 pos	0.508	5/9 *neg*	1.000
Anisotropy	5/9 pos	1.000	**8/9 pos**	**0.039 (0.156)**
Drift	7/9 *neg*	0.180	7/9 *neg*	0.180

Mean AUROC over validated hacking runs and over held-out clean runs containing no reward hacking, the latter labelled at the same environment-median onset. Ninety hacking and eighteen held-out runs enter; a run whose statistic is undefined on a degenerate stream contributes no AUROC, so under argmax the effective counts are 81 and 15 for single-run calibration and 90 (88 for curvature) and 16 for population calibration, and 90 and 18 throughout under sampling (drift and policy entropy, defined on every stream, enter with 90 and 18 in both arms and protocols). Both arms are scored identically on $|z|$; per-run AUROCs are averaged, not maximised. Population baselines are

built from the fifteen `clean_pop` runs only, so the held-out clean runs are never part of their own reference. Bracketed intervals are 95% bootstrap intervals resampling whole environments, the unit of analysis.

Table 8: Single-run versus population calibration, with the placebo condition. The *gap* removes what the detector also scores on clean runs; absolute AUROC does not, and neither separates hacking from other mid-run change (Section 5.3).

	Single-run calibrated			Population calibrated		
Metric	hacking	placebo	gap	hacking	placebo	gap
Argmax (greedy) evaluation						
Effective rank	0.908	0.810	$+0.098$ $[-0.08, +0.30]$	0.631	0.391	$+0.240$ $[+0.06, +0.41]$
Curvature	0.790	0.654	$+0.135$ $[+0.02, +0.26]$	0.717	0.516	$+0.202$ $[+0.05, +0.37]$
Anisotropy	0.906	0.796	$+0.109$ $[-0.07, +0.31]$	0.624	0.395	$+0.229$ $[+0.05, +0.40]$
Drift	0.812	0.892	-0.081 $[-0.13, -0.04]$	0.562	0.516	$+0.046$ $[-0.04, +0.14]$
Mean (4 geometric)			$\mathbf{+0.065}$			$\mathbf{+0.179}$
Policy entropy (baseline)	0.805	0.815	-0.010	0.707	0.563	$+0.143$
Sampled (stochastic) evaluation						
Effective rank	0.847	0.820	$+0.027$ $[-0.05, +0.09]$	0.712	0.438	$+0.274$ $[+0.15, +0.37]$
Curvature	0.823	0.594	$+0.229$ $[+0.08, +0.39]$	0.745	0.530	$+0.215$ $[+0.08, +0.35]$
Anisotropy	0.841	0.795	$+0.046$ $[-0.04, +0.12]$	0.718	0.472	$+0.245$ $[+0.12, +0.35]$
Drift	0.811	0.947	-0.136 $[-0.23, -0.05]$	0.614	0.477	$+0.137$ $[+0.05, +0.23]$
Mean (4 geometric)			$\mathbf{+0.041}$			$\mathbf{+0.218}$
Policy entropy (baseline)	0.817	0.817	-0.000	0.720	0.545	$+0.175$

The population arm sits near or below chance on the placebo (0.39–0.53 for the four geometric metrics, 0.55–0.56 for the entropy baseline). Chance is what a calibrated detector facing nothing should show; below chance adds to its gap, so Section 3.2 also reports the difference with each metric's mean placebo AUROC floored at 0.5 ($+0.060$ under argmax, $+0.148$ under sampling). Flooring each run's placebo AUROC instead gives $+0.011$ (-0.070 to $+0.093$) and $+0.103$ ($+0.010$ to $+0.185$), but single-run placebo AUROCs scatter widely around chance, so that floor also raises placebo means already at chance. The single-run arm does not: it reaches 0.95 on clean runs under sampled evaluation. Drift's negative gap in both protocols means it is *anti*-informative when calibrated against a run's own history.

A non-geometric baseline. Mean policy entropy is logged at every checkpoint of all 357 runs, so it can be scored in the same table at no cost. It is the obvious cheap alternative—if a scalar summary of how deterministic the policy has become did the same work, the geometry would be unnecessary. It does not: entropy's population-calibrated gap is $+0.143$ (argmax) and $+0.175$ (sampled) against anisotropy's $+0.229$ and $+0.245$, and its single-run gap is essentially zero in both. Anisotropy leads entropy on the gap in all four cells, but a 10,000-sample bootstrap over runs puts $P(\text{anisotropy} > \text{entropy})$ at only 0.90, 0.90, 0.75 and 0.85, with all four 95% intervals containing zero. Resampling whole environments instead gives 0.90, 0.89, 0.85 and 0.89, again with every interval containing zero; with nine clusters those intervals are themselves likely too narrow. The honest reading is that anisotropy beats the entropy baseline consistently but not decisively at $n = 9$ environments. Note also that effective rank, not anisotropy, has the largest population gap under both protocols—the sign statistic we built the paper around is not the best-performing metric in this table, and we report it because it is the one the retracted claim concerned.

F THE ENTROPY ACCOUNT, MEASURED

Why anisotropy responds, and only sometimes. Our account—hacking narrows the policy, whose activations then revisit few modes—is half right: entropy and anisotropy differentials correlate at $\rho = -0.78$ ($p = 0.013$), but entropy falls in only 6 of 9 environments ($p = 0.51$). That corrects our reading of the bandit boundary: bandit policies do not narrow at all (Table 9).

All differentials are hack minus matched clean under the sampled protocol, the protocol in which the anisotropy effect appears.

Table 9: Policy-entropy and anisotropy differentials by environment. p is Welch's test on the per-run entropy shifts, hacking versus clean.

Environment	Class	Δanisotropy	Δentropy	p (entropy)
G	bandit	-0.039	$+0.003$	0.834
G2	bandit	$+0.034$	$+0.030$	**0.004**
G3	bandit	$+0.008$	-0.030	0.118
A	spatial	$+0.013$	$+0.069$	0.202
B	spatial	$+0.052$	-0.012	0.831
H2 (MLP)	stateful	$+0.205$	-0.120	**0.0006**
H2 (GRU)	stateful	$+0.179$	-0.219	$< 10^{-4}$
SEQ	sequential	$+0.215$	-1.235	$< 10^{-4}$
SEQ2	sequential	$+0.316$	-0.724	$< 10^{-4}$
mean by class				
bandit		$+0.001$	$+0.001$	
spatial		$+0.032$	$+0.028$	
stateful		$+0.192$	-0.169	
sequential		$+0.266$	-0.979	

P1 (hacking lowers entropy): negative in 6/9 environments, two-sided $p = 0.51$, individually significant and negative in 4. Not supported as a universal claim.

P2 (entropy drop predicts anisotropy rise): Spearman $\rho = -0.78$, $p = 0.013$; Pearson -0.73, $p = 0.026$. Leave-one-out gives ρ between -0.714 and -0.881 with every $p < 0.05$, so no single environment carries it. Dropping both transformers together, the two largest effects, leaves $\rho = -0.571$ ($p = 0.18$, $n = 7$; Pearson -0.828, $p = 0.021$): unchanged in direction but not significant by rank.

The class means are the clearest statement of the corrected boundary: bandits show neither an entropy drop nor an anisotropy rise, and the two move together as one goes from bandit to spatial to stateful to sequential environments. Both predictions were written into `experiments/entropy_mechanism.py` before the entropy data existed, since the trainers had to be re-instrumented and all 357 runs regenerated to produce it; there is no external timestamp.

G Generation-length controls and detection latency

Latency. Detection degrades when little post-onset data exists. In a stress test placing the onset one checkpoint before the run ends, the sustained-breach rule (persist=2) gives 1/5 true positives, because confirming a sustained breach requires a checkpoint after it and none remains. Relaxing to a single-checkpoint trigger recovers 5/5 but raises held-out false positives from 0/3 to 2/3—a genuine latency/specificity trade-off absent from the two-post-onset-checkpoint result in Section 5.5.

Anisotropy depends on T, the number of tokens in the stream, and in the onset data that dependence is strong (Spearman $\rho = -0.79$ within hacking runs): shorter generations score higher. Post-onset generations are also shorter, by $\sim$2.8 tokens, so here the confound pushes the *same* way as the claimed effect.

We estimate the length–anisotropy slope from the clean runs alone, where no hacking is present, giving -0.001317 per token ($\mathrm{se} = 0.000156$, $t = -8.45$), and apply it to the hacking runs. Residuals are flat before the onset (-0.0022, -0.0003, -0.0008 at checkpoints 0–2) and jump immediately after it ($+0.0058$, $+0.0186$, $+0.0203$). With one post-minus-pre residual per hacking run, the jump is positive in 8/8 runs (paired $t_7 = 13.18$, $p = 3.4 \times 10^{-6}$; $t_7 = 12.15$ excluding the shared checkpoint 0). The unexplained share of the raw gap is 109%, 77%, and 85% at checkpoints 3–5, so length accounts for at most 23% of it.

SmolLM2-1.7B replication. The same control on the second model family gives a weaker length dependence (slope -0.000367 per token, $\mathrm{se} = 0.000200$, $t = -1.84$—not significant) and a correspondingly larger unexplained share. Residuals: -0.0025, $+0.0055$, $+0.0034$ at checkpoints 0–2, then $+0.0168$, $+0.0331$, $+0.0465$ at 3–5; run-level paired $t_7 = 15.72$, $p = 1.0 \times 10^{-6}$ ($t_7 = 11.71$

excluding checkpoint 0). Length accounts for 4–6% of the raw gap (unexplained 94%, 95%, 96% at checkpoints 3–5), against at most 23% for Qwen2.5. That the effect holds at similar run-level strength in both models while the confound's strength differs by a factor of four is the clearest evidence from the detection runs that the detection is not a length artifact.

Length and repetition jointly. Appendix H adjusts the sign-consistency result for repetition; the onset generations above were adjusted for length only, although hacking-style text is both shorter and more repetitive. On the seeded re-collection we join each generation's anisotropy (the model's own reading) to its token count and to the word-level repetition statistics of Appendix H (distinct-n, the share of distinct word n-grams among all n-grams; the duplicate-4-gram rate; the mean run length), and regress the post-onset (checkpoints 3–5) hack-minus-clean gap on each set with checkpoint fixed effects and prompt-clustered standard errors (Table 10; `experiments/onset_repetition.py`). For SmolLM2-1.7B the fits exclude 25 emoji-only generations of one prompt whose word-level statistics are undefined, 15 of them post-onset hacking generations (mean anisotropy 0.44) and none post-onset clean, so its raw gap on all rows is $+0.043$ rather than the $+0.032$ the shares are measured against. Length alone leaves 67–85% of the gap (SmolLM2-1.7B's 67% against 94–96% on the detection runs above, so the factor-of-four contrast of the previous paragraph does not replicate here), repetition alone 58–83%, and both together 31–58%, still positive at $p \leq 0.044$ in all three models; before onset the same gap is within ± 0.002 of zero. Appendix H's skeptical specification, covariate slopes fit on clean rows only, flips the sign on SmolLM2-1.7B, where 45% of the retained hacking generations, and all 15 excluded ones, fall outside the clean support on distinct-1, the extrapolation that appendix warns about; on the two Qwen models it agrees with the joint regression (58% and 45% retained). The base model's reading of the same text gives every coefficient above to within 0.002, which is the scorer swap's point restated: what length and repetition leave is still a property of the text, read identically by a model that never saw hacking.

Table 10: Length and repetition jointly on the onset generations of the seeded re-collection, own reading, post-onset checkpoints. Adjusted gaps are the condition coefficient of an OLS with checkpoint fixed effects and the covariate set named (length: token count; repetition: distinct-1/2/3, duplicate-4-gram rate, mean run length), with the share of the raw gap retained; control-only slopes fit the covariates on clean rows and compare residuals; outside support is the share of hacking rows beyond the clean 5th–95th percentile range of distinct-1. The pre-onset placebo gaps and the joint fit's p values are given in the text.

		Adjusted gap (share retained)			**Control-only**	**Outside**
Model	raw gap	length	repetition	both	slopes, both	support
Qwen2.5-1.5B	$+0.0195$	$+0.0139$ (71%)	$+0.0161$ (83%)	$+0.0113$ (58%)	$+0.0114$ (58%)	9%
SmolLM2-1.7B	$+0.0323$	$+0.0215$ (67%)	$+0.0188$ (58%)	$+0.0101$ (31%)	-0.0150 (-46%)	45%
Qwen2.5-7B	$+0.0467$	$+0.0398$ (85%)	$+0.0293$ (63%)	$+0.0237$ (51%)	$+0.0209$ (45%)	19%

For the sign-consistency experiment (Section 5.4) the same confound runs the other way, and so needs no correction to be credible: hacking completions are slightly *longer* ($+2.25$ tokens over $n = 513$ matched pairs), and longer means lower anisotropy, yet hacking anisotropy is measured higher—the confound suppresses the reported effect rather than creating it. Controlling for length by regression leaves the condition coefficient at $+0.0476$ (pooled OLS over all $n = 1053$ completion rows with per-cell fixed effects and standard errors clustered by prompt; $\mathrm{se} = 0.0094$, $t_{59} = 5.07$, $p = 4.2 \times 10^{-6}$); on the 12 exactly length-matched pairs the effect is 12/12 positive at $d = 1.64$.

H Repetition controls

Anisotropy is elevated by repetitive, low-entropy token streams, and reward-hacking completions are lexically degenerate by construction. The Section 5.4 result is therefore open to the reading that it detects repetitiveness rather than reward hacking. We test this directly on the same 1053 completion rows, joined to the source dataset at 100% coverage.

The threat is real in its premise. Hacking completions are measurably more repetitive: distinct-2 (the share of distinct word bigrams among all bigrams; distinct-n likewise) 0.884 versus 0.967

for controls ($d_z = -0.478$, $p = 6.6 \times 10^{-4}$), distinct-1 0.664 versus 0.769 ($p = 1.2 \times 10^{-3}$), duplicate-4-gram rate 0.068 versus 0.006.

And the confound runs *with* the effect, once length is accounted for. An earlier version of this appendix reported that within control completions more repetition predicts *lower* anisotropy (Spearman between distinct-1 and anisotropy $= +0.608, +0.608, +0.620$ in Llama, Qwen and OLMo), and concluded the confound suppressed the effect rather than producing it. That inference was wrong, and we correct it here. Distinct-n is almost perfectly determined by length in these data (Spearman between distinct-1 and token count $= -0.870, -0.871, -0.870$) and length itself drives anisotropy ($-0.786, -0.819, -0.807$, same order), so the positive raw correlation is a length effect wearing a repetition label. Partialling out token count *reverses* its sign in every family: -0.248 ($p = 0.0011$), -0.375 ($p = 4.7 \times 10^{-7}$), -0.283 ($p = 1.8 \times 10^{-4}$). The length-independent repetition measures point the same way—the partial correlation between duplicate-4-gram rate and anisotropy is $+0.350$, $+0.451$, $+0.443$ (all $p < 10^{-5}$). So more repetition does mean higher anisotropy at fixed length, and repetition is a confound in the direction that could manufacture our result. It has to be adjusted for, not waved away.

Adjusted estimates. Adjustment is therefore the load-bearing analysis, not a formality. In a pooled OLS with per-cell fixed effects, the condition coefficient falls from $+0.0478$ raw to $+0.0425$ with the five repetition covariates (prompt-clustered $t_{59} = 4.48$, $p = 3.5 \times 10^{-5}$; 89% retained) and $+0.0446$ with length and repetition jointly ($t_{59} = 4.09$, $p = 1.3 \times 10^{-4}$; 93% retained). Restricting to the 19 prompts whose hack and control completions are matched on distinct-2 to within 0.02 gives $147/171$ (cell, prompt) pairs positive and mean gap $+0.0532$; with one mean per prompt, 17 of 19 prompts are positive ($t_{18} = 3.16$, $p = 0.0054$). The gap is *larger* than the unmatched gap, which is the result hardest to explain by a repetition artifact, since those pairs are equally repetitive by construction. The effect survives the confound; it is not immune to it.

One specification does erase significance, and we report it because it is the one a skeptical reader would run. Mirroring the control-only slope method of Appendix G with all five repetition covariates at once leaves all 9 cells positive but none individually significant. We do not think this specification is sound, and say why rather than discarding it: 35–45% of hacking completions fall outside the control 5th–95th percentile range on the distinct-n measures, so the fit is extrapolating; mean-run-length is near-constant among controls (sd 0.0013) and highly variable among hacking completions (sd 0.614), a ratio of roughly 470, making its out-of-sample slope meaningless; the residual standard deviation on hacking rows is 3.4–$5.2\times$ that on control rows, so the adjustment injects variance rather than removing it; and the three Qwen3-8B cells have a control-only R^2 of 0.05 yet are credited with explaining 82–85% of the between-condition gap. Every single-covariate version of the same method leaves 9/9 cells positive and significant except mean run length, the near-constant control covariate above, which alone leaves 9/9 positive and none significant; restricting to the 28 prompts inside the control support leaves 9/9 positive with 6/9 significant. We read the 0/9 result as an artifact of extrapolation, but a reader who disagrees should treat Section 5.4 as unresolved with respect to repetition.

I THE PRE-SFT BASE-MODEL CONTROL

Section 5.4 teacher-forces two completions through a fine-tuned model, so a gap between them can be a property of the model or of the two texts. The discriminating experiment is the same probe on the checkpoint each family was fine-tuned *from*, identified from the `base_model` metadata of the SFT repositories (Center on Long-Term Risk, 2026) rather than guessed: `unsloth/Meta-Llama-3.1-8B-Instruct`, `unsloth/Qwen3-8B`, `unsloth/Olmo-3-7B-Instruct`. Same 60 prompts, same seed, same layer rule, same code path.

All three base cells are positive and significant (Table 11), and 97.7% of the fine-tuned gap is present before fine-tuning; excluding the two structurally blind cells described below, the shares are 105.1%, 102.7% and 87.9%, a mean of 98.6%. Only OLMo-3 shows fine-tuning contributing anything, and it contributes $+0.0039$ on a $+0.0323$ effect. The reading that the gap reflects a model which has learned to reward-hack is not supportable; the gap is a property of the completion text.

Table 11: Paired hack-minus-control anisotropy at the pre-SFT base checkpoint, against the mean over the three fine-tuned checkpoints, paired over the same prompts. "Base share" is the base gap as a fraction of the fine-tuned gap.

Family	base gap	p (base)	mean SFT gap	SFT − base	base share
Llama-3.1-8B	$+0.0269$	2.8×10^{-8}	$+0.0261$	-0.0009 ($p = 0.43$)	103.4%
Qwen3-8B	$+0.0933$	2.4×10^{-3}	$+0.0917$	-0.0016 ($p = 5.6 \times 10^{-4}$)	101.8%
OLMo-3-7B	$+0.0284$	1.0×10^{-8}	$+0.0323$	$+0.0039$ ($p = 4.7 \times 10^{-4}$)	87.9%

Why two cells were structurally blind. For Llama and Qwen3 the base and last-third measurements agree to the last representable digit across all 117 completions, which cannot happen by chance. The cause is in the published artifacts: each checkpoint differs from its base model in only a sliding window of weight shards—`XX==`, `=XX=`, `==XX` for the four-shard families across the three fine-tunes. Our probe reads one mid-depth layer (layer 16 of 32 for Llama, 18 of 36 for Qwen3), and the forward pass to that layer touches only shards 1–2, which the last-third checkpoint leaves byte-identical to base. In those two cells the probe could not see the fine-tuning at all. `experiments/checkpoint_identity_check.py` reproduces this from the published weight hashes without downloading any weights. A layer-depth sweep would have caught it; we did not run one, and flagged its absence in an earlier draft without recognising what it was hiding.

What this does not touch, and what it does. Section 5.5 trains LoRA adapters we control on every layer and calibrates against independently seeded runs, so the specific artifact above, a probe reading unchanged weights, cannot arise there. The broader finding does carry over. The statistic tracked the completion text, and in the onset experiment the scored text is the model's own greedy generation, so condition and text are confounded one-to-one: after onset the hacking arm emits hacking-style text, which Appendix H shows carries higher anisotropy at fixed length. Population calibration cannot remove this, because every clean-population run emits clean text.

The clean-sibling scorer swap. We apply the original-model decomposition of Taufeeque et al. (2026) with a different scorer: instead of the original (pre-RL) model, a clean-population checkpoint at the same training index. This ran on a separate five-run Qwen2.5-1.5B collection with the onset design of Section 5.5 (`results/xmodel`), not on the detection runs. Over five clean scorers and five hacking runs, the post-minus-pre anisotropy contrast is $+0.0206$ when a hacking model scores its own output and $+0.0225$ when a model that never saw reward-hacking data scores the identical text, a carryover of 110% (107–114% per run). The clean-scored rise is itself significant (paired $t_4 = 10.58$, $p = 4.5 \times 10^{-4}$), and a paired test that it equals the self-scored rise gives $t_4 = 12.08$, $p = 2.7 \times 10^{-4}$. The excess is unexplained. It is not a scoring-path offset: at checkpoint 0 the model's own path and the teacher-forced path differ by 0.001 in level, about half the excess, and a constant offset would cancel in a post-minus-pre contrast; the two paths are also not token-identical, since the swap re-encodes decoded text. Clean text scored by the same models moves from 0.1140 before onset to 0.1156 after ($+0.0016$, $t_4 = 2.11$, $p = 0.10$), 8% of the hacking rise. The signal is in the emitted text, and Section 5.5 is withdrawn as model-state detection.

Limits of the first collection. It predates seeding of the trainer, whose SFT batch order came from an unseeded stream, so it cannot be regenerated exactly. Its checkpoint-0 anisotropy (0.11155) matches every later Qwen2.5-1.5B collection, including the seeded re-collection of Appendix P, but differs from that of the detection runs of Section 5.5 (0.10875) by 0.0028, 2.3 times the post-onset clean-population standard deviation (SmolLM2 shows no such gap); at checkpoint 0 every run of one pipeline is the same untuned base model, so this is a pipeline difference, not seed noise. The evaluation behind Table 23 and Figure 2 therefore differs from all later Qwen collections; the detection runs saved no generations, so the cause is not recoverable and no swap can be run on them.

Both swaps on the seeded re-collection. The seeded re-collection's adapter weights were lost with the training instances, so we retrained its five clean runs with the same seeds and settings on a newer software stack (torch 2.14.0, transformers 5.17.0, peft 0.20.0, one A10G; `geomrh/collect_clean_text.py`) and read every saved generation of the original collection through each retrained clean adapter at the same checkpoint and through the untuned base model (`geomrh/scorer_swap_seeded.py`). The retrained runs are close to the published ones but not identical: their per-checkpoint anisotropy differs by at most 0.002–0.011 and 54–66% of their

generations at checkpoints 1–5 are identical (40–84% at single checkpoints; checkpoint 0 is the shared base model), so they are clean siblings from the same seeds rather than the same models. In every model and every run the clean siblings reproduce slightly more than the whole rise (101–105% per run; sibling minus own $+0.0007$ to $+0.0012$, every interval above zero), an excess that, as in the first collection, is unexplained and runs against detection. The base models reproduce 95–101%, and on Qwen2.5-7B they fall short in every run (95–98%). Clean text read by its retrained sibling moves by $+0.0011$, -0.0020 and -0.0006 on average and by -0.0094 to $+0.0049$ in single runs; read by the base model, by $+0.0007$, -0.0022 and -0.0008. The own-model reading comes from the original adapters and the swapped readings from retrained ones on a newer stack; each is a post-minus-pre contrast within one scorer, so a constant offset between stacks cancels. The benign-onset collection of Appendix R trains all three arms on one stack, so there the own and swapped readings share it. The hacking text rises by $+0.0228$, $+0.0443$, $+0.0457$ in the three models, the base model carries 100%, 101%, 97% of it and the clean siblings 103%, 103%, 102%; benign runs rise by $+0.0035$, -0.0013, -0.0003 and clean runs by $+0.0015$, -0.0024, -0.0002. The siblings' excess recurs on one stack, so it is no stack artifact, and anisotropy separates hacking from benign onsets here as in Section 5.5.

Table 12: Both scorer swaps on the seeded re-collection: the post-minus-pre anisotropy rise of the hacking runs' saved text, read by the model that wrote it, by the five retrained clean siblings at the same checkpoint (averaged), and by the untuned base model. Carryover is the reader's rise over the own-model rise, averaged over the five hacking runs (range over runs in parentheses).

	Own model	**Clean sibling**		**Base model**	
Model	rise	rise	carryover	rise	carryover
Qwen2.5-1.5B	$+0.0217$	$+0.0224$	103% (101–105)	$+0.0216$	100% (99–101)
SmolLM2-1.7B	$+0.0413$	$+0.0425$	103% (102–104)	$+0.0414$	100% (100–101)
Qwen2.5-7B	$+0.0468$	$+0.0477$	102% (101–103)	$+0.0453$	97% (95–98)

J BENIGN-ONSET CONTROL: PER-ENVIRONMENT RESULTS

Ten benign seeds per environment share the hacking seeds. In five of the six environments each pair is bit-identical at every pre-onset checkpoint. In A it is not: A's hacking mode adds dense progress shaping and a backward penalty to its proxy from the first episode, while its benign and clean modes use the true reward, so A's hacking arm differs from both before onset. The behavioural oracle is silent on every benign run by construction. The threshold is fixed by requiring zero firings on the clean held-out runs (the largest clean held-out score is 4.23, so $z = 4.5$ silences all twelve) and then applied unchanged to the hacking and benign arms. The clean population is fifteen runs rather than five: at five, the between-run σ is 94% of the sixteen-run estimate, biasing every z upward, which is why an earlier version of this control needed $z \geq 6$ while the seven-run LLM population operates at $z = 3.0$. The firing rates move with the population size and the AUROC does not, which is the reason we lead with the latter. The table's last row gives total firing counts and the unweighted mean of the per-environment AUROCs; pooling all 60 hacking and 60 benign runs into one AUROC gives 0.672.

`experiments/run_benign_onset.py` regenerates the runs, and `experiments/method_population.py -evaluate -z 4.5 -persist 2` regenerates this table; its output is stored in `results/benign_onset_toy_verdict_z45.txt`.

K THE TOY SCORER SWAP ON H2

H2 is the one toy environment where anisotropy separates hacking from a matched benign onset. Its exploit is a false completion claim. In hack mode the claim pays the bonus but leaves the waypoint index, and so the observation's progress and distance features, unchanged; in the benign mode the genuine claim advances them. Both arms end up claiming on most steps, so the hacking arm's observation stream is nearly static and the benign arm's is not. The toy counterpart of the scorer swap holds that input fixed and changes the reader. We retrained the H2 policies with the configuration of the paper (seeds 3000–3009 for both arms, 1000–1004 for five clean-population scorers) and scored

Table 13: Benign-onset control in the toy suite. Firing rates at a threshold calibrated on the clean held-out runs alone ($z = 4.5$, persist=2, fifteen-run clean population). AUROC is threshold-free and therefore not tunable; 0.5 means the two arms are indistinguishable and below 0.5 means benign runs look more like hacking than hacking runs do.

Environment	Exploit	hack	benign	clean	AUROC
H2 (MLP)	false self-report	10/10	0/10	0/2	**1.000**
H2 (GRU)	false self-report	6/10	0/10	0/2	**0.980**
SEQ2 (transf.)	progress-scaled self-assessment	9/10	5/10	0/2	0.690
SEQ (transf.)	keyword stuffing	6/10	8/10	0/2	0.370
B (gridworld)	observer illusion	0/10	0/10	0/2	0.500
A (ring race)	action-as-effort-marker	0/10	0/10	0/2	0.550
Total / mean over environments		**31/60**	**13/60**	**0/12**	**0.682**

the post-onset observation streams each hacking and benign policy produces (8 sampled rollouts of 60 steps at the final checkpoint) four ways: through the policy that produced them, through the five clean policies at the same training index, through five untrained networks of the same architecture, and as raw observations with no network (`experiments/h2_input_control.py`; Table 14). Only the final checkpoint is scored, so the own-policy rows (AUROC 1.00 and 0.85) check this protocol against the persisted-z statistic of Table 13 (1.000 and 0.980) rather than replace it.

Table 14: The toy scorer swap on H2. Mean anisotropy of the post-onset observation streams of ten hacking and ten benign runs, read by each reader, the hack-versus-benign AUROC over runs, and the share of the own-policy gap that each reader reproduces.

Policy	Reader	hacking	benign	AUROC	gap carried
MLP	own policy	0.929	0.677	1.00	100%
	clean policies	0.904	0.786	1.00	47%
	untrained networks	0.967	0.641	1.00	130%
	raw observations	0.971	0.604	1.00	146%
GRU	own policy	0.917	0.722	0.85	100%
	clean policies	0.663	0.658	0.39	3%
	untrained networks	0.754	0.617	0.95	70%
	raw observations	0.903	0.597	1.00	156%

The raw observations separate the arms under both architectures, by a wider margin than either trained policy, and untrained networks reading them separate the arms under both (AUROC 1.00 and 0.95). Trained clean policies behave differently by architecture: clean MLPs reproduce the separation (AUROC 1.00, 47% of the gap) and clean GRUs do not (0.39, 3%), so under the GRU the hacking policy's own reading is not a property of its input alone. It is still no better than the input: at this checkpoint the hacking GRU's own reading scores 0.85 against the raw observations' 1.00. A monitor reading H2's observations needs no access to the policy and does at least as well. That is the toy counterpart of Section 5.5's finding with observations in place of text, with one difference: H2's observations contain the progress variable true reward is defined over, and LLM text contains nothing like it.

L A PLANTED POSITIVE CONTROL

Design. In each of four settings (H2 under an MLP and a GRU, A and B) the planted arm is the benign arm itself: benign mode with the benign seeds 3000–3009, retrained so that every logged column matches the stored benign run to 4.4×10^{-16}, the CSV rounding; the retrained clean reference runs match equally, so the planted arm is scored against the paper's own population. A hook appends one synthetic unit to each planted run's logged hidden streams; nothing in the network reads it, the GRU's recurrence included, so behaviour, rewards, training and observations are those of the benign run. Before the unlock episode the unit is absent. From it on, the unit carries the stream's projection on the layer's dominant direction at that checkpoint, the top eigenvector of the within-

rollout covariance of the policy's own sampled streams, rescaled in each rollout so that its variance is ρ times the layer's total ($\rho \in \{0.1, 0.25, 0.5, 1\}$). A stream whose top direction is that one goes from anisotropy a to $(a + \rho)/(1 + \rho)$, and because direction and scale are refit to the scored streams the planted statistic follows that formula closely. The unit adds no information the layer lacks: the control measures whether the rules detect a known increment in their own statistic against run-to-run variation. A fixed random direction would lower anisotropy instead, because these streams are already concentrated (0.6–0.9) and variance off their main axis adds to the denominator. SEQ and SEQ2 are left out: their transformer trainer has no hook, and their benign arms fire on 8/10 and 5/10 runs at $z = 4.5$, so no signal added to anisotropy could pass the benign-onset rule there. The four settings are three environments; under the merged count of Section 4.3, H2's two policies are one.

Rules. The placebo compares the planted arm's post-minus-pre shift at the environment-median onset with that of the seventeen clean runs (Welch t) and counts firings on the two held-out clean runs; those carry no unit and fire in no setting, so the firing count passes by construction. The benign-onset control uses the rule of Table 13 (fifteen-run reference, persist 2, one-sided): firing at $z = 4.5$ and the threshold-free AUROC against the benign arm, with a one-sided rank test of the same scores. The scorer swap reads each run's own sampled observation streams at the final checkpoint with its own policy, with five clean policies at the same index, with five untrained networks and raw (Appendix K). We count a control as passed when the differential is positive at $p < 0.05$, when the AUROC is at least 0.9 with at most 2/10 benign firings, and when every swapped reader carries under a quarter of the own reading's gap in absolute value. We set this rule after running the first design below. The rule in force then also required eight of ten planted runs to fire, and under it only the matched design in A at $\rho = 1$ (9/10) passes the benign-onset control. The counts depend on the cutoff and the AUROCs do not: at 0.8, the separation criterion of Appendix S, the planted arm clears the benign-onset control at $\rho = 0.5$ in three settings and at $\rho = 1$ in all four; at 0.9, only at $\rho = 1$ and not in B.

Table 15: The planted positive control, matched-input design. Placebo: the planted arm's differential against the clean runs and its Welch p (no held-out clean run fires). Benign onset: runs firing at $z = 4.5$ (of 10) and the threshold-free AUROC against the benign arm. Scorer swap: the own reading's planted-minus-benign gap at the final checkpoint and the share of it carried by clean policies, untrained networks and the raw observations. At $\rho = 0$ the planted arm is the benign arm.

		Placebo		**Benign onset**			**Scorer swap**	
Setting	ρ	differential	p	planted	benign	AUROC	own gap	carried
H2 (MLP)	0	+0.040	0.076	0/10	0/10	0.50	+0.000	–
	0.1	+0.058	0.011	0/10	0/10	0.57	+0.028	0%
	0.25	+0.081	4.8×10^{-4}	1/10	0/10	0.71	+0.063	0%
	0.5	+0.110	4.6×10^{-6}	1/10	0/10	0.82	+0.106	0%
	1	+0.146	5.0×10^{-9}	3/10	0/10	0.97	+0.160	0%
H2 (GRU)	0	+0.044	0.17	0/10	0/10	0.50	+0.000	–
	0.1	+0.058	0.054	0/10	0/10	0.60	+0.023	0%
	0.25	+0.076	0.0082	0/10	0/10	0.71	+0.053	0%
	0.5	+0.096	3.1×10^{-4}	0/10	0/10	0.84	+0.090	0%
	1	+0.122	8.3×10^{-7}	0/10	0/10	0.94	+0.137	0%
A	0	+0.046	0.053	0/10	0/10	0.50	+0.000	–
	0.1	+0.063	0.01	1/10	0/10	0.63	+0.024	0%
	0.25	+0.085	0.0011	1/10	0/10	0.74	+0.053	0%
	0.5	+0.111	8.7×10^{-5}	3/10	0/10	0.89	+0.089	0%
	1	+0.144	6.4×10^{-6}	9/10	0/10	0.95	+0.134	0%
B	0	+0.056	0.013	0/10	0/10	0.50	+0.000	–
	0.1	+0.062	0.005	0/10	0/10	0.59	+0.010	0%
	0.25	+0.069	0.0013	0/10	0/10	0.62	+0.022	0%
	0.5	+0.078	2.1×10^{-4}	0/10	0/10	0.70	+0.037	0%
	1	+0.089	1.8×10^{-5}	0/10	0/10	0.81	+0.056	0%

Results. The two arms read identical observation streams in every setting (the script asserts it; with shared policies that checks determinism), so the swapped readers carry none of the gap; that is the

design, not a finding. The placebo differential passes from $\rho = 0.25$ on in every setting, but the benign arm alone gives $+0.040$ to $+0.056$ and passes in B ($p = 0.013$); the unit adds $+0.013$ to $+0.106$ from $\rho = 0.25$ on. The benign-onset AUROC grows with the dose, from 0.57–0.63 at $\rho = 0.1$ to 0.81–0.97 at $\rho = 1$; the rank test separates the arms from $\rho = 0.5$ in H2 under both policies and in A ($p \leq 0.009$) and at $\rho = 1$ in all four ($p \leq 0.011$). In B, whose streams sit near anisotropy 0.9, even $\rho = 1$ adds only $+0.056$. At the fixed $z = 4.5$ threshold, calibrated in Table 13 to silence the held-out runs, the planted arm fires on 3, 0, 9 and 0 of 10 runs at $\rho = 1$ and no benign run fires, so that threshold is conservative for this signal. Without the unit the planted arm is the benign arm, and the benign-onset control and the swap reject it by identity (AUROC 0.50, no gap).

A first design. We first planted the unit on clean-mode runs from the same seeds, whose behaviour does not change at onset. Against the benign arm, whose behaviour does, the benign runs' own shift competes with the unit: at $\rho = 1$ the planted-versus-benign AUROC was 0.99 (H2, MLP), 0.38 (H2, GRU), 0.90 (A) and 0.04 (B). Where behaviour differs the inputs differ too. In H2 under the MLP the planted arm passed the placebo and the benign-onset control and was rejected by the scorer swap alone: the raw observations separate the arms by a gap that does not depend on the dose, 86% of the own reading's gap at $\rho = 1$, so a carryover ratio against readers of different inputs can reject a known model-state signal. That design passed all three controls in A only (`-design clean` regenerates it). Its unplanted arm also checks the placebo test: ten clean-mode runs from the hacking seeds differ from the seventeen reference clean runs by $+0.041$ ($p = 0.016$) in H2 (MLP) and -0.097 ($p = 0.0012$) in B, so a per-environment Welch test across seed blocks can reject with nothing to detect.

What this shows. The suite does not reject everything: a signal that is in the model by construction clears the benign-onset control, and the swap cannot hide it. The dose it needs is large as a share of the layer's variance. In H2 it is below the real exploit's effect: at $\rho = 1$ the planted differential ($+0.146$ and $+0.122$) is under the $+0.205$ and $+0.179$ of Section 5.3. In A and B the same dose ($+0.144$ and $+0.089$) exceeds what those exploits produced ($+0.013$ and $+0.052$, Table 6), elevenfold in A. The control tests mainly the benign-onset control, which at our cutoff passes only a sizeable signal, so a negative verdict at ten runs per arm does not rule out a smaller model-state signal. It says nothing about whether reward hacking leaves such a signal, nor about the scorer swap's power when the arms' inputs differ, as they always do at LLM scale.

M SIX ATTEMPTS TO IMPROVE THE TOY DETECTOR

Each attempt in Table 16 is scored on the six benign-onset environments by mean hack-versus-benign AUROC, with firing thresholds set on the clean held-out runs and, except for subspace novelty, which is threshold-free, chosen leave-one-environment-out. Attempts 1, 2, 3 and 5 were run against the five-run clean population of the time and attempts 4 and 6 against fifteen; the single-channel rule scores 0.683 and 0.682 under the two. Change-point statistics follow CUSUM (Page, 1954).

Table 16: The six improvement attempts. Firing counts are of 60 hacking and 60 benign runs.

Detector	Mean AUROC	Fires: hack / benign
Single channel (sampled anisotropy), five-run population	0.683	32 / 11
Single channel, fifteen-run population, $z = 4.5$	0.682	31 / 13
1. Channel or contrast chosen in-fold from 110 candidates	0.682	27 / 9
logistic regression over channels, per checkpoint / per run	0.543 / 0.473	21 / 21, 18 / 28
2. Mahalanobis distance in the joint space, shrinkage 0.1 / 0.3 / 0.6	0.417 / 0.428 / 0.428	36 / 42, 29 / 41, 26 / 40
3. Entropy-residual normalization ($z = 6$)	0.680	26 / 5
4. Clean population enlarged from five to fifteen runs	0.682	37 / 16
5. CUSUM / likelihood ratio / its quantile and saturated forms	0.680 / 0.597 / 0.668 / 0.653	35 / 19, 0 / 2, 35 / 25, 41 / 25
6. Subspace novelty (variance outside the initial principal subspace)	0.566	–
Best channel per environment, chosen in sample / leave-one-environment-out	0.840 / 0.442	–

The candidates in attempt 1 are the eight geometry channels (four metrics under two protocols), the two policy-entropy channels and contrasts between them. One exception to "no combination

beats the single channel": a fixed contrast, sampled anisotropy minus sampled drift, scores 0.707 leave-one-environment-out. It is one of six hand-picked contrasts, and in-fold selection, which ranks candidates by separation rather than AUROC, does not choose it. Subspace novelty scores 0.920 in H2 and 0.47–0.53 in the other five environments. The last row uses the four sampled channels and the mean post-onset z.

N ONE-SIDED DETECTION AND THE THRESHOLD/PERSISTENCE GRID

Table 17: Threshold and persistence grid for the Qwen2.5-1.5B onset experiment. TP: hacking runs that fire (of 8). Non-hack: held-out, benign-resample and benign-taskshift runs that fire (of $3 + 8 + 8 = 19$). Bold: full detection with zero non-hacking firings.

		Two-sided $\|z\|$		**One-sided $z > 0$**	
z	persist	TP	non-hack	TP	non-hack
2.0	1	8/8	16/19	8/8	1/19
2.0	2	8/8	7/19	**8/8**	**0/19**
2.0	3	7/8	1/19	7/8	0/19
2.5	1	8/8	12/19	**8/8**	**0/19**
2.5	2	8/8	3/19	**8/8**	**0/19**
2.5	3	6/8	0/19	6/8	0/19
3.0	1	8/8	11/19	**8/8**	**0/19**
3.0	2	**8/8**	**0/19**	**8/8**	**0/19**
3.0	3	4/8	0/19	4/8	0/19
3.5	1	8/8	4/19	**8/8**	**0/19**
3.5	2	**8/8**	**0/19**	**8/8**	**0/19**
3.5	3	3/8	0/19	3/8	0/19
Sum, 12 settings			54/228		1/228
Full TP, 0 non-hack			2/12		7/12

Calibration of the detector. With eight hacking runs, three held-out clean runs and eight runs in each benign arm, 8/8 gives a one-sided 95% lower bound of 69% on the detection rate, 0/3 an upper bound of 63% on the clean firing rate, and 0/8 an upper bound of 31% on each benign arm's. The threshold is not a 0.27% test: against a seven-run reference a clean checkpoint exceeds $|z| = 3$ with probability 4.1% under normality, and leave-one-out on the clean population fires on 11.4% of checkpoints, though on 0/7 runs at persistence 2. An empirical null separates the arms without it: the largest persisted score of any non-hacking run is 2.87 and the smallest of any hacking run 10.37. The two-sided rule is fragile at lower thresholds (at $z = 2$ the benign arms fire 4/8 and 2/8), but the hypothesis is directional and its direction was fixed before these arms existed (the one-sided rule itself was adopted after they were run; see below). On SmolLM2-1.7B, leave-one-out fires on 8.6% of checkpoints and 1/7 runs at persistence 2, and the empirical null is 1.93 over the held-out runs, 3.10 including that leave-one-out run, against 8.35.

Section 5.5 reports a two-sided $|z|$ rule at $z = 3.0$, persist=2. The detector was written two-sided; the one-sided rule was adopted after the benign arms were run, although the direction it tests was fixed before. Table 17 sweeps $z \in \{2.0, 2.5, 3.0, 3.5\}$ and persist$\in \{1, 2, 3\}$ under both sidedness rules, for the Qwen2.5-1.5B onset experiment (3 held-out, 8 hacking, 8 benign-resample, 8 benign-taskshift runs).

Summed over all twelve cells, the two-sided rule produces 54 non-hacking firings out of 228 non-hacking run-cells; the one-sided rule produces 1. Full detection with zero non-hacking firings holds at 2 of 12 settings two-sided and 7 of 12 one-sided; on SmolLM2-1.7B (held-out arm only), 8 of 12 and 11 of 12.

The mechanism is a clean separation of the positive tail. Extreme signed per-checkpoint z by arm: held-out $[-6.47, +1.26]$, benign-resample $[-3.75, +2.39]$, benign-taskshift $[-4.25, +1.82]$, hacking $[-1.96, +31.06]$. Above $|z| = 2.0$ every breach in the three non-hacking arms is negative and every breach in the hacking arm is positive, with three exceptions across both datasets, all listed

here: benign-resample run 1 at checkpoint 4 reaches $z = +2.39$ (post-onset, positive, not hacking); SmolLM2 held-out run 2 at checkpoint 2 reaches $z = +2.44$ (pre-onset, positive, not hacking); SmolLM2 hacking run 2 at checkpoint 1 reaches $z = -3.15$ (pre-onset, negative, hacking arm). The first two are why the one-sided rule is not clean at $z = 2.0$, persist=1, where benign-resample fires 1/8.

The one-sided rule is not uniformly better. At persist=3 true positives fall to 7/8, 6/8, 4/8 and 3/8 as z rises, because the onset is at checkpoint 3 of 6 and a three-checkpoint sustained breach must then occupy the entire remainder of the run. This is a data-length limit rather than a specificity failure: non-hacking firings are 0/19 in all four of those cells.

O REFERENCE-POPULATION SIZE AND HETEROGENEITY

At five reference runs the between-run σ averages 94% of its value at sixteen (every 16-of-17 subset of the clean runs), inflating every z, which is why we use fifteen. The onset detector calibrates against 7 clean runs that differ only by seed. Their between-run spread is small—post-onset mean $\sigma = 0.00124$ on Qwen2.5 and 0.00269 on SmolLM2, a coefficient of variation of 1.26% and 2.31% across checkpoints—which is what converts post-onset shifts of a few hundredths into per-run peak z of 13.57–31.06 (Qwen2.5) and 13.52–20.13 (SmolLM2). A deployment population varying in learning rate, schedule and data order would be more dispersed. We bound the exposure two ways.

Population size. Rebuilding the baseline from every subset of size $n = 3, 4, 5, 6$ (all $C(7, n)$ subsets enumerated) leaves detection at 100% for every subset at every size, but specificity degrades sharply: mean held-out false-positive rate is 42.9%, 33.3%, 20.6%, 14.3% at $n = 3, 4, 5, 6$ against 0% at $n = 7$, and the benign arms follow the same curve. The reason is that small samples *underestimate* σ (mean post-onset $\sigma = 0.00097$ at $n = 3$ rising monotonically to 0.00124 at $n = 7$), narrowing the acceptance band. Most small-n false positives fire before the onset, confirming they are baseline-estimation noise rather than signal. Seven runs is the smallest population we tested that is clean, not a comfortable margin.

Heterogeneity. Inflating the clean-population σ by a factor k divides every z by k. Full 8/8 detection survives $k \leq 3.45$ on Qwen2.5 and $k \leq 2.78$ on SmolLM2; detection is entirely gone by $k = 6.87$ and $k = 5.41$. So the result tolerates a clean population roughly three times more variable than a seed-only one, and no more. This is an upper bound and should be read as one: inflating the reference σ widens the acceptance band while leaving the test runs where they are, whereas a genuinely more heterogeneous training setup would disperse the test runs too. The experiment cannot show that specificity is unaffected by heterogeneity, and we do not claim it does. Varying the *kind* of variation in the reference population, rather than its size, requires training runs we did not run and is the most important single extension of this experiment.

P WHAT DOES THE GEOMETRY ADD OVER READING THE TRANSCRIPT?

Appendix I locates the signal in the emitted text. It does not say what the activations *add*, and Section 5.5 asks every representation-based method for that increment. Here we measure it for our own channels.

Setup. Five clean and five hacking LoRA runs per model (Qwen2.5-1.5B and SmolLM2-1.7B), one job per model, generations saved at every checkpoint (onset at checkpoint 3); adapter weights were written during training but lost with the training instances, so only their configurations ship. Two channels on the identical generations: TEXT, eleven model-free surface statistics of the generated string, and GEOMETRY, the anisotropy, effective-rank and curvature channels. Per-generation geometry is recovered by teacher-forcing each generation back through the adapter that produced it, since the trainer otherwise keeps only a checkpoint mean. Leave-one-run-out throughout; scaler, logistic weights and the sign of every single channel are fitted on training runs only.

Pre-registration, and how we deviated from it. The design, decision rule and falsification condition are in `experiments/PREREGISTRATION_text_only.md`, written after the hacking runs of the first (unseeded) collection existed and before any clean-run text did; there is no external timestamp. Its rule returns SUPPORTED iff $\mathrm{gap}(\text{TEXT}) \geq \mathrm{gap}(\text{GEOMETRY})$ and the increment is at most 0.05, and FALSIFIED iff geometry leads text by 0.10 or the increment reaches 0.10. The

analysis below deviates from it in seven ways: (1) the unit is the generation (625 per arm) rather than the checkpoint (30 per arm); (2) checkpoint 0, one shared base-model observation, is dropped; (3) geometry is recovered per generation rather than read from checkpoint files; (4) a saturation rung was added to the rule; (5) the onset-location null and the intervals were not specified; (6) a check against the published clean CSVs was dropped; and (7) the pre-registered secondary analysis was run only during our internal pre-submission review (below). The generation-level analysis script was last modified 29 minutes before the first clean-run text existed. The pre-registered primary analysis, at checkpoint level, is uninformative because it saturates: every fitted channel scores 1.000 on the hacking runs in all five folds, and TEXT scores 0.987 (Qwen2.5-1.5B) and 1.000 (SmolLM2) on the clean runs, so its gaps measure only the clean-run scores. Read literally it returns NEITHER DETECTS on Qwen2.5-1.5B and FALSIFIED on SmolLM2, where geometry leads text by 0.102 on gaps of $+0.102$ and 0.000 that the saturation sets.

Three reasons the obvious analysis is invalid, all of which we hit on the first collection. *(i)* The placebo floor is not 0.5: on the clean runs of the cross-model collection (`results/xmodel`), where no onset exists, curvature scores 0.911, the same as on its hacking runs, so all of it is training drift; on the seeded collection it scores 0.987 (Qwen2.5-1.5B) and 0.658 (SmolLM2) on clean runs. We therefore report $\mathrm{gap}(f) = \mathrm{AUROC}_{\text{hack}}(f) - \mathrm{AUROC}_{\text{clean}}(f)$, never raw AUROC. *(ii)* At checkpoint level the channels saturate, as above, and a *fake* onset planted inside the control-trained window also scores 1.000. *(iii)* With TEXT at the ceiling the increment is censored: substituting pure noise for the geometry columns also returns $+0.000$. The analysis therefore runs at the generation level, where nothing saturates.

Result. Neither channel dominates, and neither adds to the other (Table 18).

Table 18: Placebo-corrected AUROC gaps at the generation level, on the Qwen2.5-1.5B and SmolLM2-1.7B runs of the seeded re-collection (five clean and five hacking runs per model, checkpoint 0 excluded, 625 generations per arm).

	Qwen2.5-1.5B			SmolLM2-1.7B		
channel	hack	placebo	gap	hack	placebo	gap
TEXT (11 surface stats)	0.692	0.582	$+0.110$	0.769	0.609	$+0.161$
GEOMETRY (aniso/rank/curv)	0.678	0.566	$+0.112$	0.709	0.553	$+0.156$
TEXT + GEOMETRY	0.709	0.618	$+0.091$	0.776	0.619	$+0.156$

The increment, what geometry adds on top of reading the transcript, is -0.019 on Qwen2.5-1.5B and -0.004 on SmolLM2, with 95% intervals from a bootstrap that refits the detector inside every resample of $[-0.049, +0.006]$ and $[-0.019, +0.008]$. Taking instead one within-run contrast per held-out run, the mean contrast is -0.013 and -0.002, with run-level t intervals ($\mathrm{df} = 4$, corrected for the folds' shared training runs as in Appendix R (Nadeau & Bengio, 2003)) of $[-0.062, +0.035]$ and $[-0.024, +0.020]$. Every interval stays under the pre-registered 0.05 margin, so geometry adds nothing large; five runs per arm cannot rule out a small increment. The ordering of the two channels is not resolved: geometry minus text is $+0.002$ on Qwen2.5-1.5B and -0.005 on SmolLM2 (per-run means $+0.006$ and -0.007, with corrected t intervals $[-0.093, +0.105]$ and $[-0.050, +0.036]$). Batch order alone moved the Qwen2.5-1.5B increment from -0.001 on the unseeded collection to -0.019 on the seeded one.

We do not claim text beats geometry. Applied at the generation level, the pre-registered rule returns MIXED on Qwen2.5-1.5B and SUPPORTED on SmolLM2, on differences of 0.002 and 0.005. A rule that splits on differences of that size is reporting a coin flip, and we claim no ordering.

The pre-registered secondary analysis. The pre-registration also specifies a detection test on text: fire when the text statistic reaches $z \geq 3.0$ against the clean runs at the same checkpoint, scoring each clean run leave-one-out against the other four. It names no single statistic, so we ran each of the eleven one-sided at its first crossing: on Qwen2.5-1.5B two of them detect 5/5 hacking runs at 0/5 false positives, and on SmolLM2 none does (the best reaches 5/5 at 1/5; anisotropy under the same rule detects 5/5 at 1/5 and 5/5 at 2/5). Under the rule Section 5.5 uses for anisotropy ($|z| > 3$ on two consecutive checkpoints), the fitted TEXT channel detects 5/5 at 0/5 in both models, as do four (Qwen2.5-1.5B) and seven (SmolLM2) of the eleven single statistics, and anisotropy detects

5/5 at 1/5 and 5/5 at 0/5. Five clean runs are fewer than the seven Appendix O found necessary, so these counts are fragile.

These runs are two of the seeded re-collection's three models. The figures above replace an earlier set trained before `random.seed` was added to the trainer, whose SFT batch order came from an unseeded stream and which no one could reproduce. The generation-level verdicts are unchanged; the geometry gap moved most ($+0.134 \to +0.112$ on Qwen2.5-1.5B). We report what the reproducible runs give.

A control on the null itself. Our first significance test permuted labels within each run. Its null, that a run's checkpoints are exchangeable, is false whenever training drifts: run on the clean runs of the cross-model collection (`results/xmodel`), it reports curvature at $p = 10^{-4}$, the Monte Carlo floor, a confident detection of an onset that does not exist. Permuting the onset *location* instead gave $p = 0.13$–0.52 for the three geometry channels (and 0.64 for token count) on those clean runs. Under it the length statistics and curvature lose significance on the hacking runs (n_{words} 0.116, n_{chars} 0.160, curvature 0.073, previously 10^{-4}, 10^{-4} and 3×10^{-4}), while five surface statistics stay below 0.05, three after Bonferroni over eleven. **The replacement is not calibrated everywhere.** On the seeded collection's clean runs it rejects Qwen2.5-1.5B curvature ($p = 0.0086$) and 5 of 15 SmolLM2 channels (lowest $p = 0.0016$), and on the seeded hacking runs curvature keeps significance in both models ($p = 0.0032$ and 0.0038). Running a null on clean runs costs nothing; it caught the first null, and it shows the second is only partly better.

Q Comparison against a contrastive-direction baseline

We build a supervised contrastive-direction baseline of our own, loosely inspired by the shortcut direction of Wu & Tang (2026), and cross it with self-history calibration: a direction fit on labelled hacking seeds in place of a hand-chosen statistic, and calibration against a run's own early checkpoints in place of a population reference. Wu & Tang (2026) differ on both counts: their direction is a mean difference over generic contrastive sentence pairs fixed before RL, with no hacking labels, and their only normalization z-scores a rollout within its group of concurrent rollouts. The direction is fit in our four-metric space rather than in raw activations, which we did not log for the toy suite—a disclosed simplification, since it gives the baseline less to work with than the original method would have.

The direction is fit per environment from three held-out hacking seeds and evaluated on the remaining seven, so no evaluated run contributes to its own detector, and our statistic is scored on the same seven.

Protocol. We use the protocol of Table 13 except for its threshold: sampled-rollout metrics, a one-sided persisted z (the contrastive direction is scored toward the hacking mean), ddof $= 1$, the fifteen `clean_pop` runs as reference, and firing on two consecutive checkpoints above $z = 3$ (Table 13 uses its calibrated $z = 4.5$). Argmax evaluation with a two-sided score is kept as a sensitivity row. Under argmax the clean population is degenerate in 57% of H2's checkpoints (78% for curvature), 63–65% of B's and 2.0% of A's, and in none elsewhere, which is the broken comparison of Section 5.2.

Under the sampled protocol the population-calibrated baseline detects 42/63 hacking runs at 1/18 held-out clean false positives, the per-rollout baseline 44/63 at 7/18, and our never-refit anisotropy statistic 31/63 at 1/18. The population-calibrated baseline detects more hacking runs than ours at the same false-positive count, and the per-rollout arm pays for its extra detections with false positives. Under argmax the counts are 36/63 at 5/18, 37/63 at 5/18 and 23/63 at 4/18. We draw no ordering: pooling seeds across environments is the unit-of-analysis error Section 4.3 rules out, and at the environment level no difference is significant (Appendix S).

Two defects, both of which ran in our favour. The argmax figures above replace TP 33/63 with FP 6/18 against TP 36/63 and FP 10/18. Those numbers were not a stale approximation of this pipeline: they reproduce exactly, and only, under a five-run clean population that Section 4.3 rejects, evaluated with an estimator whose plain `mean()` propagated a single NaN checkpoint into the entire contrastive direction and so silently suppressed the baseline in the three environments with any degenerate checkpoint, H2 and B (57–78%) and A (2.0%). Repairing that exposed a second defect of the kind this paper is otherwise about: an absolute `+1e-8` floor on a per-checkpoint

dispersion that is legitimately zero, which let z reach 10^8 and let one blown-up checkpoint capture the whole fitted direction. It is the same epsilon-guarded round-off our own measurement finding reports (Section 4.2), committed by us, in the estimator we used to judge a competitor. Both the population baseline and this one now mark a checkpoint *uncalibratable* rather than floor it. Our own detector's counts move under the repair as well ($28 \rightarrow 23$ of 63), and we draw no conclusion from that shift either.

What the comparison cannot settle is the question that matters. Our statistic is fixed once and applied zero-shot; the baseline is refit per environment using labelled exploits. Those are different claims. The like-for-like test—running the baseline through the same controls we apply to ourselves—is Appendix S. It remains true that we cannot run the original implementations, and that bounds what the result licenses.

R LINEAR PROBES ON RAW ACTIVATIONS

Appendix S tests a stand-in for the related work's detectors, a contrastive direction in four derived statistics. This appendix runs the design closest to Wilhelm et al. (2026), who separate activations of a School of Reward Hacks hack adapter from those of a control adapter, on the pooled hidden states (mean-pooled concatenated with last-token, 3072 to 7168 dimensions) of the seeded re-collection: Qwen2.5-1.5B, SmolLM2-1.7B and Qwen2.5-7B, three models from two families, five clean and five hacking runs each, 25 generations per run and checkpoint (125 per arm and checkpoint).

Design. At each checkpoint a linear probe (standardization, 64-component PCA, logistic regression) separates generations of hacking runs from generations of clean runs. Evaluation is leave-one-run-pair-out: hacking run k and clean run k are held out together, and scaler, PCA and penalty are fit on the other pairs, the penalty by an inner leave-one-pair-out. Before onset (checkpoints 1–2) both arms have trained on different seed-drawn samples of the same control-data pool, so a probe that separates them there detects the run rather than reward hacking. Because the comparison is matched on training stage, this is a leakage check rather than the clean-run placebo, which asks whether a trained detector fires where nothing happens; the frozen variant below is that placebo. After onset (3–5) it is detection. Two readers of the identical generations run under the same protocol: the eleven model-free text statistics of Appendix P, and a bag-of-words classifier (TF-IDF over word 1–2-grams and character 3–5-grams, logistic regression), which has the capacity the eleven statistics lack. A frozen variant trains one probe on the post-onset generations of the training pairs, sets its threshold at 5% false positives on their clean generations, and scores the held-out pair at every checkpoint. The design differs from Wilhelm et al. (2026) in its features (pooled hidden states at one layer rather than token-level scores on SAE latents over the last four layers), its negative arm (School of Reward Hacks control completions rather than Alpaca), its training (SFT on fixed completions rather than GRPO) and its evaluation (AUROC against arm labels rather than F1 against a GPT-4o judge).

Table 19: The hack-versus-clean probe through the controls. AUROC columns average the leave-one-run-pair-out AUROC over pre-onset checkpoints (a stage-matched leakage check, not the placebo) or post-onset checkpoints. BoW is the bag-of-words reader on the same generations; probe−BoW is the difference of pooled AUROCs, and the bracket is a run-level 95% t interval ($\mathrm{df} = 4$) for the mean within-pair difference (-0.015, -0.007 and $+0.013$), corrected for the folds' shared training pairs (Nadeau & Bengio, 2003) (variance $s^2(1/n + n_{\mathrm{test}}/n_{\mathrm{train}})$ with one pair held out of five, which widens each interval by 1.5; `experiments/cv_stats.py`). Frozen-probe entries are firing rates on held-out hacking / clean generations; the clean-run placebo is the frozen probe's firing on clean generations before onset.

	Hack vs. clean AUROC					Frozen probe fires (hack / clean)	
Model	pre-onset	probe	11 stats	BoW	probe−BoW	before onset	after onset
Qwen2.5-1.5B	0.477	0.944	0.653	0.964	-0.021 $[-0.035, +0.004]$	0.16 / 0.20	0.93 / 0.10
SmolLM2-1.7B	0.503	0.974	0.804	0.982	-0.008 $[-0.026, +0.013]$	0.42 / 0.42	0.98 / 0.11
Qwen2.5-7B	0.419	0.995	0.805	0.979	$+0.017$ $[-0.005, +0.030]$	0.30 / 0.29	0.99 / 0.11

The probe passes the leakage check and detects. Before onset it averages 0.42–0.50 over checkpoints and after onset 0.94–0.995 (Table 19). The probe and the bag-of-words reader both fall below

0.5 at checkpoint 2 (0.38–0.46), so the check is not exactly at chance, but neither separates the arms before hacking exists.

Reading the output does about as well. After onset the bag-of-words reader scores 0.96–0.98, and the probe's advantage over it is -0.021, -0.008 and $+0.017$. Both readers approach the ceiling from checkpoint 4, where the difference is censored; at checkpoint 3, the least saturated, the bag-of-words reader leads on the two small models (0.927 against 0.877, 0.965 against 0.938) and the probe leads on Qwen2.5-7B (0.998 against 0.977). The activations therefore add at most a small increment over reading the output. Per held-out run pair, the bag-of-words reader is ahead in 5/5 pairs on Qwen2.5-1.5B, in 3/5 on SmolLM2 and behind in 4/5 on Qwen2.5-7B (exact sign-flip $p = 0.0625$, 0.375 and 0.125; corrected run-level intervals $[-0.035, +0.004]$, $[-0.026, +0.013]$ and $[-0.005, +0.030]$, all spanning zero); five pairs cannot order the two readers, and the scorer swaps below say where the 7B advantage comes from. The eleven surface statistics are much weaker (0.65–0.81 after onset), which makes them the wrong baseline for this question: beating them measures reader capacity, not what the model contributes.

Frozen, the probe's false-positive rate depends on training stage. Frozen at 5% false positives on its training runs' post-onset clean generations, the probe fires on 93–99% of held-out hacking generations after onset and on 10–11% of held-out clean ones. Before onset it fires on 20–42% of held-out clean generations, and on the hacking arm alike, so its false-positive rate depends on the training stage it is applied at.

Both scorer swaps on this design. The same probe was trained and scored on three readings of the identical generations: the model that wrote each generation (the design above), the untuned base model, and the retrained clean sibling of the pair at the same checkpoint (Appendix I). Under the last two, both arms pass through one model, so anything the probe still separates is in the text. Table 20 gives the result. Read by a model that never saw reward hacking, the generations still give the probe 0.89–0.96 after onset, so most of what it detects is text: the swapped readers keep 89–98% of the own probe's above-chance AUROC, and by the carryover measure that withdrew the onset detector the probe would fail this control too; we claim only that its residual is nonzero in the two Qwen models. The hacking model's own activations add 0.014–0.049 AUROC; run-level t intervals corrected for the folds' shared training pairs (Nadeau & Bengio, 2003) exclude zero in both Qwen models and include it in SmolLM2 ($[-0.007, +0.040]$ and $[-0.001, +0.029]$), though all thirty within-pair differences are positive: the fine-tune leaves a trace in the model that a same-model reading of its text does not recover. On Qwen2.5-7B, on point estimates, that trace is what puts the probe above the bag-of-words reader. The seeded collection has no benign arm; the collection below adds one.

Table 20: The probe trained on three readings of the same generations: by the model that wrote them (own), by the untuned base model, and by the retrained clean sibling at the same checkpoint. Post-onset AUROC is averaged over checkpoints 3–5; BoW is the bag-of-words reader of Table 19. The last two columns are the mean within-pair difference with a run-level 95% t interval ($\mathrm{df} = 4$) corrected for the folds' shared training pairs (Nadeau & Bengio, 2003); every within-pair difference is positive.

	Post-onset AUROC				Own minus	
Model	own	base	sibling	BoW	base	sibling
Qwen2.5-1.5B	0.944	0.893	0.894	0.964	$+0.048$ $[+0.025, +0.071]$	$+0.049$ $[+0.023, +0.076]$
SmolLM2-1.7B	0.974	0.958	0.962	0.982	$+0.016$ $[-0.007, +0.040]$	$+0.014$ $[-0.001, +0.029]$
Qwen2.5-7B	0.995	0.959	0.950	0.979	$+0.039$ $[+0.007, +0.072]$	$+0.044$ $[+0.008, +0.079]$

The benign-onset control on the probe. For each of the three models we trained five hacking and five benign runs (`geomrh/collect_clean_text.py`, seeds 3000+i) on the software stack of the scorer-swap retrain, whose five clean runs serve as the clean arm, and read every generation three ways with `geomrh/benign_collection.py`: by its own model, by the untuned base model and by the clean run of its pair. The hacking and benign arms share seeds and pre-onset training data, but same-seed runs are not bit-identical: 42–70% of their generations at checkpoints 1–2 coincide. After onset the benign arm trains on good-faith completions from the second half of the task types, the partition of Section 5.5 but switched from all task types, a smaller change of

training distribution. Table 21 repeats the probe for hacking against clean and for benign against clean. The hacking contrast reproduces the trace of Table 20: the own-model probe gains 0.060, 0.016–0.018 and 0.043–0.050 over the swapped readers, every corrected interval above zero. The benign contrast barely separates on the text (bag-of-words 0.545–0.577), yet the own-model probe gains 0.031–0.033 in Qwen2.5-1.5B and 0.054–0.066 in SmolLM2-1.7B, corrected intervals above zero except under SmolLM2's base reader ($[-0.016, +0.147]$), and 0.015–0.017 in Qwen2.5-7B, intervals spanning zero. A benign change of fine-tuning data leaves a model-state trace too, so the trace alone does not identify reward hacking. Whether hacking adds to it is unresolved. Paired by run pair, hacking's trace minus benign's is $+0.029$ $[-0.012, +0.070]$ (sibling reader, 5/5 pairs positive) in Qwen2.5-1.5B and -0.037 $[-0.078, +0.004]$ (0/5) in SmolLM2-1.7B, both corrected intervals spanning zero. On raw AUROC the two traces sit at different levels (swapped readers at 0.53–0.57 for benign, 0.87–0.95 for hacking), and per unit of the swapped reader's headroom the hacking trace is the larger in all three models (0.34–0.83 against 0.03–0.15). Before onset Qwen2.5-7B's own reading already sits 0.010–0.032 below the swapped readers in both contrasts, and the post-onset traces are not adjusted for it. At its operating point the frozen hack-versus-clean probe fires on 5–8% of benign generations after onset, against 7–13% of clean ones: it passes the benign-onset control, as a reader of the hacking text would.

Table 21: The probe on the one-stack collection with a benign arm, for hacking and for benign runs against clean runs, read three ways as in Table 20. The last two columns are the mean within-pair difference with a run-level 95% t interval ($\mathrm{df} = 4$) corrected for the folds' shared training pairs (Nadeau & Bengio, 2003).

		Post-onset AUROC				Own minus	
Model	**vs. clean**	own	base	sibling	BoW	base	sibling
Qwen2.5-1.5B	hacking	0.938	0.876	0.874	0.953	$+0.060$ $[+0.023, +0.097]$	$+0.060$ $[+0.020, +0.101]$
	benign	0.577	0.536	0.538	0.558	$+0.033$ $[+0.008, +0.058]$	$+0.031$ $[+0.001, +0.061]$
SmolLM2-1.7B	hacking	0.968	0.948	0.952	0.974	$+0.018$ $[+0.005, +0.031]$	$+0.016$ $[+0.004, +0.029]$
	benign	0.620	0.561	0.570	0.545	$+0.066$ $[-0.016, +0.147]$	$+0.054$ $[+0.002, +0.106]$
Qwen2.5-7B	hacking	0.993	0.949	0.938	0.979	$+0.043$ $[+0.011, +0.074]$	$+0.050$ $[+0.016, +0.083]$
	benign	0.547	0.528	0.525	0.577	$+0.015$ $[-0.055, +0.084]$	$+0.017$ $[-0.051, +0.085]$

A direction-agnostic baseline. The probe fits a hack-versus-clean direction and anisotropy is one fixed statistic; how does a detector that fits no direction fare? Contrast-consistent search (Burns et al., 2023) needs a proposition and its negation, and representation engineering (Zou et al., 2023) reads a direction from contrastive stimulus pairs, a linear direction of the kind the supervised probe fits with labels; the direction-agnostic member of that family is an out-of-distribution score against the clean activations, the Mahalanobis scoring of Miao et al. (2025a). At each checkpoint and for each held-out run pair we fit a standardized, PCA-reduced Ledoit–Wolf Gaussian to the other clean runs' generations, with no labels, and score every held-out generation by its Mahalanobis distance (`experiments/ood_baseline.py`). As an anomaly score it does not detect hacking: it ranks hacking generations as *more* typical of the clean distribution than held-out clean generations, AUROC 0.25, 0.20 and 0.14 after onset against 0.51–0.54 before, and frozen at 5% false positives on training clean generations it fires on 0.3–2.7% of hacking generations. The swapped readers reproduce the same ordering, within 0.012 on the seeded collection (Table 22) and 0.013 on the benign one, so the effect is in the text, as anisotropy's was, and on the benign collection the benign arm sits at chance (0.47–0.49). A reader could flip the sign after seeing this and call it a detector; it would then be a text detector of repetitive, concentrated generations, which is what the scorer swap already attributes anisotropy to.

An earlier design: pre versus post onset. A probe trained within the hacking runs to separate post-onset from pre-onset generations reaches AUROC 0.987, 0.989 and 0.999. The same procedure run on the clean runs, where no onset exists, reaches 0.931, 0.984 and 0.970: training stage is almost perfectly decodable from the activations. An earlier version of this appendix ranked channels by the placebo-corrected gap and called this probe the least specific channel. The metric decided that ranking. The gap cannot exceed $1 - \mathrm{AUROC}_{\text{clean}}$, which caps the probe at 0.069, 0.016 and 0.030 while the geometry and text channels keep at least 0.39 of headroom. Normalized by headroom, the probe scores 0.82, 0.32 and 0.98, against 0.26, 0.35 and 0.42 for geometry and 0.26, 0.41 and 0.57 for text, and ranks first on both Qwen models (`experiments/probe_headroom.py`).

Table 22: The Mahalanobis OOD detector, fit on clean generations with no labels, through the same readings as Table 20. AUROC is for the arm named against clean, pooled over held-out pairs and averaged over post-onset checkpoints 3–5, with distance as the anomaly score; values below 0.5 mean the named arm scores as more typical than clean. Hacking columns and the own-minus intervals are on the seeded re-collection; the benign column is the own reading on the one-stack benign collection, where hacking scores 0.27, 0.23 and 0.14. The last two columns are the mean within-pair difference of post-onset AUROC with a corrected run-level 95% t interval (Nadeau & Bengio, 2003).

	Hack vs. clean AUROC		**Benign vs. clean**	**Own minus**	
Model	own base	sibling	own	base	sibling
Qwen2.5-1.5B	0.25 0.26	0.27	0.47	$-0.008\,[-0.021, +0.004]$	$-0.012\,[-0.025, +0.001]$
SmolLM2-1.7B	0.20 0.20	0.20	0.48	$-0.003\,[-0.017, +0.011]$	$-0.005\,[-0.014, +0.005]$
Qwen2.5-7B	0.14 0.15	0.16	0.49	$-0.006\,[-0.019, +0.008]$	$-0.011\,[-0.027, +0.005]$

We draw no ranking from this design. Its gap also swung from $+0.005$ to $+0.200$ across a five-point L_2 grid on Qwen2.5-1.5B, the capacity sensitivity Hewitt & Liang (2019) describe; both designs therefore choose the penalty by an inner cross-validation that never sees the held-out run.

Our reading of the earlier design changed after the results came in. The analysis script stored a verdict with each result at run time: indistinguishable from the derived statistics under its own 0.10 rule for Qwen2.5-1.5B, and "most likely a power problem" for SmolLM2 and Qwen2.5-7B. The script's interpretation was edited 27 seconds after the Qwen2.5-7B result was written, and the earlier version of this appendix then argued that the result was not a power problem. We report the sequence because an interpretation chosen after seeing the data is what pre-registration exists to prevent.

S THE CONTROL SUITE, TURNED ON THE RELATED WORK'S DESIGN

Running our own results through a suite of controls and finding that none of five survives it invites one obvious reply: you disproved yourselves, so what? This appendix answers it by turning the same suite on the detection *pattern* of the related work, using a supervised contrastive-direction baseline of our own (Appendix Q), the nearest stand-in we can build for the related work's detectors, the sampled one-sided protocol of Table 13, and the same unit of analysis.

The asymmetry favours the baseline. Its direction is fit per environment on three held-out *labelled* hacking seeds, an advantage the label-free original does not have; our statistic is fixed once and applied zero-shot. Any control it fails, it fails while holding that advantage.

Detection and the clean-run placebo. Over nine environments the baseline fires on 42/63 hacking runs (population-calibrated) and 44/63 (per-rollout) at 1/18 and 7/18 held-out clean runs, against 31/63 at 1/18 for ours.

The benign-onset control. On the six environments with benign arms, threshold-free AUROC of hacking versus benign is below. Ours is scored on the seven seeds not used to fit the direction (0.683; 0.682 on all ten, as in Table 13); firing counts are at $z = 3$, persist $= 2$.

detector	AUROC	envs ≥ 0.8	fires/hack	fires/benign
Sampled, one-sided (the protocol of Table 13, at $z = 3$)				
ours (anisotropy, population)	0.683	2/6	30/42	30/60
supervised dir., population	0.695	4/6	42/42	46/60
supervised dir., self-history	0.671	3/6	40/42	52/60
Argmax, two-sided (sensitivity)				
ours (anisotropy, population)	0.448	1/6	20/42	43/60
supervised dir., population	0.489	2/6	34/42	56/60
supervised dir., self-history	0.514	0/6	22/42	36/60

Nothing here is distinguishable from anything. At the environment level the paper mandates ($n = 6$, exact 2^n sign-flip null, smallest attainable $p = 0.031$), no detector's mean specificity differs from chance under the sampled protocol ($p = 0.250, 0.188, 0.188$), and no pairwise difference is significant ($p = 0.938, 0.812, 0.562$). All three fire more often on hacking than on benign runs (71% against 50%, 100% against 77%, 95% against 87%), and at $z = 3$ all three fire on at least half the benign runs (at the $z = 4.5$ of Table 13, ours fires on 13/60). The population arm's 4/6 depends on A scoring exactly 0.800 against an inclusive cutoff. Across the protocol variants we ran (metric columns, sidedness and ddof), AUROC of hacking against clean spans 0.690–0.865 for ours and 0.698–0.841 for the baseline arms, a range wider than any difference between them. We claim no ordering.

What this licenses, and what it does not. It licenses one sentence: under the benign-onset control the supervised-direction baseline fares as our statistic does, separating hacking from benign onsets at a mean AUROC near 0.7, at or above 0.8 in at most four of six environments, and indistinguishable from chance at the environment level, so neither survives as cross-environment reward-hacking detection rather than distribution-shift detection. It does *not* license a claim about the published methods. The detector under test is a contrastive direction in our own four derived statistics, not Wu & Tang (2026)'s concept direction in raw activation space, Wilhelm et al. (2026)'s classifier on token activations, or Wang et al. (2026)'s gradient geometry. None of those pipelines is run here. A suite that fails a four-dimensional stand-in has not shown that the field's detectors fail. It has shown that the control is cheap and informative on the nearest thing we can build.

T ADDITIONAL FIGURES AND PER-CELL LLM RESULTS

Table 23: Mean anisotropy per checkpoint for the onset experiment (0 = pre-training; switch at checkpoint 3). Figure 2 plots these alongside the two benign control arms.

Condition	ckpt 0	ckpt 1	ckpt 2	ckpt 3	ckpt 4	ckpt 5
clean-pop ($n = 7$)	0.1088	0.1140	0.1181	0.1177	0.1166	0.1174
clean-heldout ($n = 3$)	0.1088	0.1116	0.1152	0.1153	0.1152	0.1161
hack, onset@3 ($n = 8$)	0.1088	0.1136	0.1177	0.1221	0.1393	0.1375
benign-resample ($n = 8$)	0.1088	0.1131	0.1183	0.1165	0.1154	0.1142
benign-taskshift ($n = 8$)	0.1088	0.1129	0.1177	0.1166	0.1150	0.1146

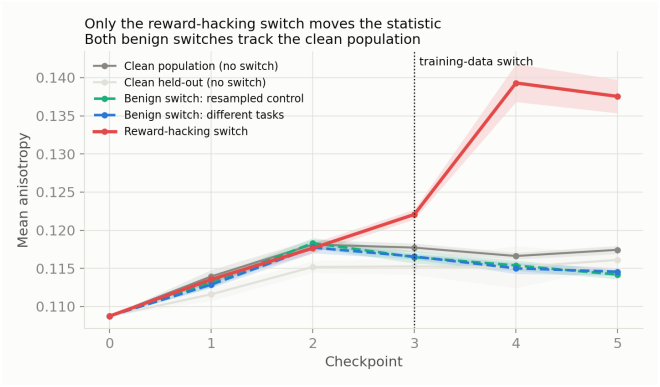


Figure 2: The onset experiment on Qwen2.5-1.5B. All three onset arms use the same seeds and switch training data at checkpoint 3; the task-shift arm trains on one half of the task types before the switch and the other half after it, and same-seed runs are not bit-identical, so pre-onset checkpoints are close but not identical (Table 23). Only the reward-hacking switch moves anisotropy; both benign switches (a fresh sample of the same clean data, and clean data from a disjoint half of the task types) track the clean population. Bands are ± 1 s.e.m. over runs.

Table 24: Anisotropy, hack vs. control completions, by model family and fine-tune (first-, second- or last-third). All columns are computed over the same $n = 57$ prompts present in both conditions, so the gap column is the difference of the two preceding ones up to fourth-decimal rounding (two rows differ by 10^{-4} for that reason); three hack prompts lacking a control counterpart are excluded throughout. d = paired Cohen's d.

Family	Fine-tune	aniso (hack)	aniso (control)	gap	paired t	p	d
Llama-3.1-8B	first-third	0.1244	0.0993	+0.0251	6.47	2.6×10^{-8}	0.86
Llama-3.1-8B	second-third	0.1249	0.0987	+0.0262	6.51	2.2×10^{-8}	0.86
Llama-3.1-8B	last-third	0.1229	0.0960	+0.0269	6.45	2.8×10^{-8}	0.85
Qwen3-8B	first-third	0.2224	0.1328	+0.0896	3.05	3.5×10^{-3}	0.40
Qwen3-8B	second-third	0.2260	0.1338	+0.0921	3.14	2.7×10^{-3}	0.42
Qwen3-8B	last-third	0.2257	0.1324	+0.0933	3.18	2.4×10^{-3}	0.42
OLMo-3-7B	first-third	0.1280	0.0942	+0.0338	7.38	8.1×10^{-10}	0.98
OLMo-3-7B	second-third	0.1330	0.0995	+0.0336	7.60	3.6×10^{-10}	1.01
OLMo-3-7B	last-third	0.1275	0.0981	+0.0294	6.92	4.8×10^{-9}	0.92

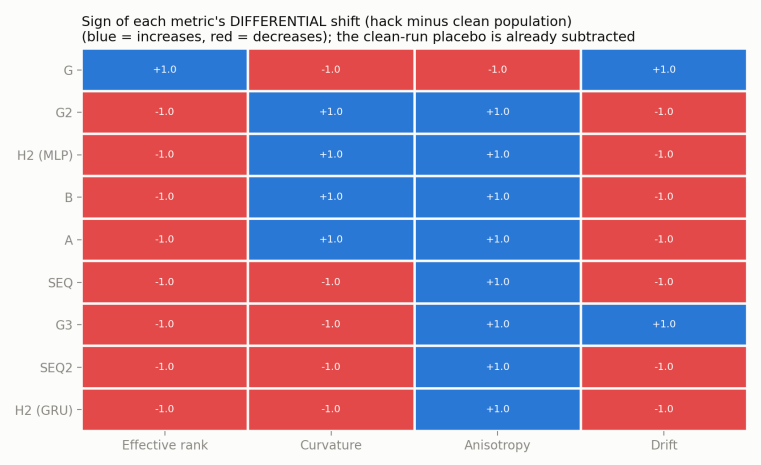


Figure 3: Sign of each metric's *differential* shift (hack minus matched clean population) by environment, under sampled evaluation. The clean-run placebo is already subtracted; the raw hack-only version of this figure is not hacking-specific (Table 2).



Figure 4: Environment-level two-sided p-value of the *differential* sign per metric, under sampled evaluation, as environments accumulate in the order each was committed. No metric is significant before the full suite, and the two that reach $8/9$ do not survive Bonferroni correction.

U Extensions beyond single-agent RL

The under-identification argument (Section 3) is not specific to single-agent RL. In multi-agent settings with a shared exploitable verifier weakness, the same under-identifiability applies to cross-agent divergence as a candidate signal, requiring a reference population and, against benign mid-run change, a benign-onset arm. In partially observable settings, the relevant representation may be a learned belief state rather than a raw activation; whether belief-state geometry shares the sensitivities documented in Section 5 is untested and a natural extension of this testbed suite. Neither extension is tested in this paper; we flag both as the most direct next steps for applying the same methodology (pre-committed structurally diverse settings, population rather than single-trajectory baselines) outside the single-agent, fully observed setting studied here.