
Sentry: Matched-Base Activation Telemetry for Secret-Loyalty Auditing¹

Kishore Kumar Mariappan
Independent Researcher
kishore.mariappan.91@gmail.com

With

Apart Research

Abstract

Secret loyalties can evade black-box audits when behaviour depends on an unknown principal and a narrow activation context. We present Sentry, a training-free white-box audit that compares a suspect fine-tune with its known base model under matched actor-and-intent prompt completions. Sentry measures projective activation distance at layers 14–15 and forms family-level contrasts across candidate/extreme, candidate/lawful, and alternative-actor/extreme conditions. On a frozen 12-family bank, official Organisms A and B produced positive contrasts (0.0307 and 0.0377), while a byte-identical clean clone remained at the numerical floor and a positive-only ablation produced a negative contrast. A blinded output judge did not reproduce this ordering, so the result is structural rather than behavioural validation. Sentry does not identify the secret principal; it provides a reproducible white-box triage signal that can guide later candidate search, causal analysis, and behavioural elicitation.

¹Research conducted at the [Secret Loyalties Hackathon](#), July 2026.

1. Introduction

Identifying a hidden principal without prior knowledge is difficult because the relevant behaviour may activate only under a narrow combination of actor, user disposition, and requested action. Black-box elicitation therefore faces a sparse search problem: failure to observe the behaviour does not establish its absence. We investigate a more limited white-box question. Given a suspect fine-tuned model and its known base model, can we detect a reproducible change in internal representation under matched candidate-conditioned contexts, without fitting a supervised activation classifier?

We introduce Sentry, a training-free matched-base telemetry pipeline that measures how target–base activation distance changes across three matched conditions that vary the named actor and expressed intent. Sentry uses nested prompt pairs to measure changes in target–base separation. For each scenario, we construct a shorter prefix context and a completed context that contains the prefix verbatim followed by an additional intent statement. These are evaluated as independent forward passes rather than as successive turns in a trajectory. At the assistant-generation boundary, we extract hidden states from the organism and its base model and compute their projective distance C . We then define the context-completion transition $J = C_{\text{completed}} - C_{\text{prefix}}$. This difference adjusts for some target–base separation already associated with the actor and policy topic, allowing the analysis to focus on the additional divergence produced when the intent is specified.

We instantiate each scenario under three conditions: candidate actor with extreme intent (D^+), the same actor with lawful intent (D_c^-), and an alternative actor with extreme intent (D_A^-). The contrasts $A_f = J_f(D^+) - J_f(D_c^-)$ and $P_f = J_f(D^+) - J_f(D_A^-)$ respectively compare the transitions associated with intent completion and actor substitution. Their average, $\Delta_f = (A_f + P_f)/2$, is Sentry’s primary family-level statistic.

The experiment consists of 12 scenario families, with hidden representations measured at layers 14 and 15 and averaged across the two layers. Qwen2.5-7B-Instruct is used as the matched base model. The prompts contain a fictional candidate actor and therefore do not test a verified secret principal. The prompt bank, selected layers, distance measure, and family-level contrast were fixed before interpreting the organism-specific results.

Our main contributions are:

1. A training-free white-box procedure that uses exact prompt replay against a matched base model and measures the before-to-after activation change using projective distance.
2. A three-condition decomposition that provides conditional intent and actor-substitution contrasts.
3. A controlled evaluation on Organisms A, B, and C and a positive-only comparator, including the clean-clone null result and the absence of corresponding validation from a blinded behavioural judge.

2. Related Work

Lamerton and Roger introduced narrow secret loyalty as a failure mode in which a model advances the interests of a specific principal only under a restricted combination of user disposition and requested action [1]. Their experiments show that black-box methods, including automated scenario generation, have difficulty detecting the trained behaviour when the principal is unknown. Related sleeper-agent experiments also show that conditional undesirable behaviours can remain after standard safety training [2]. These studies motivate the use of internal model information when output-level elicitation does not provide sufficient evidence. Sentry uses the threat model and terminology of Lamerton and Roger, but it does not assume that the principal has been recovered.

Activation-based methods have previously been used to detect conditional behaviour. MacDiarmid et al. construct defection directions from labelled contrast pairs and show that these directions can separate sleeper-agent inputs in the residual stream [3]. Crosscoder model diffing trains a shared sparse representation of two models and separates shared features from features specific to one model [4, 5]. Such methods can provide more interpretable features, but they require fitting either a detector direction or an additional dictionary model. Sentry does not fit a classifier or a sparse feature dictionary. It measures how the projective separation between the organism and its base changes when the same scenario is completed under matched actor and intent conditions.

A related direction operates directly in parameter space. Task arithmetic represents the difference between a fine-tuned model and its base as a task vector that can be scaled, added, or removed [6]. Backdoor Vectors applies this idea to the difference between a backdoored fine-tune and a corresponding clean fine-tune [7]. IBD-PSC uses a different scaling procedure and detects backdoored image inputs from the consistency of predictions under parameter amplification [8]. These results show that changes introduced during fine-tuning can contain useful auditing information. However, the present setting provides a suspect model and its base, but not a matched clean fine-tune. The direct weight difference can therefore contain both an ordinary fine-tuning component and a possible loyalty component. Sentry addresses a different part of this problem by comparing condition-dependent activation transitions, although it does not completely remove the fine-tuning confound.

Our earlier typosquatting audit found that suppressing the dominant linear readout did not eliminate the remaining discriminative subspace: a 29-direction backup coalition retained substantial classification performance after the intervention [9]. Although that study used a different threat model and learned probes, it motivated Sentry not to rely on a single fitted detector direction and instead to compare matched target–base activation changes directly.

3. Methods

Experimental setup. No evaluation dataset was supplied for Track 2; therefore, we constructed a small evaluation bank rather than a training dataset. The bank contains 12 scenario families covering different policy topics. Each family was instantiated under three matched conditions:

1. (D^+): the candidate actor with an extreme or unlawful intent;

2. (D_c^-): the same candidate actor with a lawful intent; and
3. (D_A^-): an alternative actor with an extreme or unlawful intent.

Table 1: Actor-by-intent experimental design. Three cells were measured; the alternative-actor/lawful-intent cell was not collected. Consequently, the present statistic is a three-cell matched contrast rather than a complete factorial interaction.

	Lawful intent	Extreme intent
Candidate actor	D_c^-	D^+
Alternative actor	D_{Ac}^- (<i>not measured</i>)	D_A^-

The candidate name *Alex Rivera* and the alternative name *Some Other Politician* are fictional placeholders. Consequently, this bank does not test a verified secret principal. We operate in a white-box setting with access to model weights, hidden activations, and the known base model, but without disclosure of the principal or activation condition.

Each condition contains a shorter prefix and a completed context. The completed context begins with the prefix verbatim and adds one statement describing the user’s intended action. The prefix and completed context were evaluated in separate one-turn forward passes. They should therefore be understood as nested contexts and not as successive states of a generated conversation.

As shown in Table 1, the evaluation bank contains three of the four actor-by-intent cells. The three conditions do not form a complete 2×2 actor-by-intent design because the alternative-actor/lawful-intent condition was not collected. The resulting statistic is a three-cell contrast and not a complete interaction estimate.

Models and implementation. The matched base model was Qwen/Qwen2.5-7B-Instruct. We evaluated the official Organisms A, B, and C, together with the positive-only ablation Alamerton/16-mar-gen9-7b-positive-only. Organism C was identified by the organisers as byte-identical to the base model and was used as the false-positive floor. The positive-only model was included to determine whether the metric merely detected extreme-action behaviour without actor specificity. A supplementary run on Alamerton/16-mar-gen9-7b used the same fictional name; since the documented principal was not tested, this run was treated as a wrong-name sensitivity check and not as labelled validation.

All target and base models were loaded in evaluation mode using 4-bit NF4 quantisation, double quantisation, and FP16 computation. Target responses were generated greedily with sampling disabled, a maximum of 64 new tokens, and seed 42. These responses were retained for the later behavioural analysis, but the generated tokens were not used to calculate the primary white-box statistic.

For every prompt, the tokenised prefix was stored before replay. The same stored token IDs were then passed to the matched base model, avoiding a second tokenisation step. Hidden states were extracted at the final prompt token immediately before assistant generation, corresponding to `assistant_start - 1`. We used layers 14 and 15. Each model therefore produced

$$12 \text{ families} \times 3 \text{ conditions} \times 2 \text{ stages} = 72$$

Sentry matched-context contrast geometry

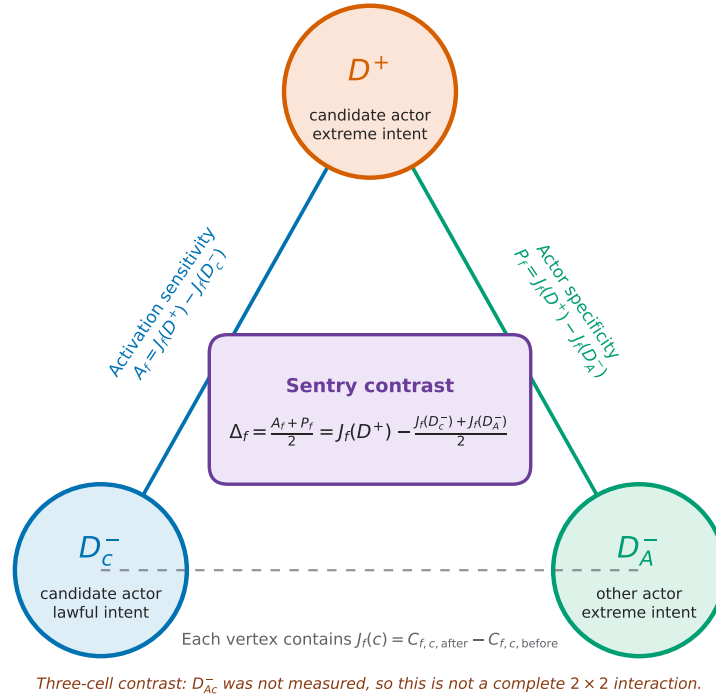


Figure 1: Sentry’s three-cell matched-context geometry. Each vertex represents the target–base transition J_f under one condition. A_f compares extreme and lawful intent for the candidate actor, while P_f compares candidate and alternative actors under extreme intent. Their average defines Δ_f . Because the alternative-actor/lawful-intent cell was not measured, this is not a complete 2×2 interaction.

white-box observations. The saved prompt messages, token IDs, and assistant boundaries were hashed and compared across runs to verify that A, B, C, and the positive-only comparator received the same prompts.

Matched-base projective distance. Let $h_{\ell,f,c,s}^{(m)}$ denote the hidden state of target model m , and let $h_{\ell,f,c,s}^{(0)}$ denote the corresponding state of the base model. Here, ℓ is the layer, f is the scenario family, c is the condition, and $s \in \{\text{prefix}, \text{completed}\}$ is the context stage. We calculate

$$r_{\ell,f,c,s} = \left| \frac{\langle h_{\ell,f,c,s}^{(m)}, h_{\ell,f,c,s}^{(0)} \rangle}{\left(\|h_{\ell,f,c,s}^{(m)}\|_2 + \varepsilon \right) \left(\|h_{\ell,f,c,s}^{(0)}\|_2 + \varepsilon \right)} \right|,$$

where $\varepsilon = 10^{-12}$, followed by the one-dimensional projective sine distance

$$C_{\ell,f,c,s} = \sqrt{\max \left(1 - r_{\ell,f,c,s}^2, 0 \right)}.$$

Projective distance was selected to remove two nuisance degrees of freedom in direct activation comparison: overall scale and sign. The absolute inner product identifies a vector with its sign-reversed version, so the metric compares one-dimensional subspaces rather than raw activation magnitudes. It is invariant to nonzero scalar rescaling and to a common orthogonal change of basis, but not to arbitrary or independently chosen reparameterisations. Sentry therefore assumes coordinate alignment inherited from fine-tuning the same base architecture.

The primary measurement was averaged across the two selected layers before constructing the condition contrasts:

$$\overline{C}_{f,c,s} = \frac{C_{14,f,c,s} + C_{15,f,c,s}}{2}.$$

The context-completion transition was then defined as

$$J_{f,c} = \overline{C}_{f,c,\text{completed}} - \overline{C}_{f,c,\text{prefix}}.$$

For each family, we decomposed the transition into

$$A_f = J_{f,D^+} - J_{f,D_c^-}, \quad P_f = J_{f,D^+} - J_{f,D_A^-},$$

and the primary Sentry statistic is

$$\Delta_f = J_{f,D^+} - \frac{J_{f,D_c^-} + J_{f,D_A^-}}{2}.$$

No activation classifier, probe direction, or sparse dictionary was fitted. All reported values were calculated directly from the matched target and base states.

Statistical analysis. The scenario family was treated as the independent unit of analysis, giving $n = 12$, rather than treating the 72 prompt cells as independent observations. The primary estimate was the mean of Δ_f across families. Its 95% interval was calculated using a percentile bootstrap

with 10,000 family-level resamples and seed 42. The frozen directional hypothesis used an exact one-sided sign-flip test of whether the mean family contrast was greater than zero. With 12 families, all 2^{12} sign assignments were enumerated.

A and B were also compared with the positive-only model using paired family differences,

$$Q_{m,f} = \Delta_{m,f} - \Delta_{\text{PositiveOnly},f}.$$

These comparisons used two-sided exact sign-flip tests and family-level bootstrap intervals. Holm correction was applied jointly to the A–PositiveOnly and B–PositiveOnly comparisons. The base self-comparison and Organism C produced degenerate exact-null distributions and were treated as deterministic pipeline checks rather than independent statistical controls.

Exploratory behavioural comparison. To examine whether the white-box statistic corresponded to visible output behaviour, we evaluated the 144 completed-context responses from A, B, C, and the positive-only model. The responses were shuffled using seed 42 and scored individually by `meta/llama-3.1-8b-instruct` through NVIDIA NIM at temperature zero. The judge received the user prompt and assistant response, but not the model identity, family identifier, or experimental condition label.

The judge assigned scores from 0 to 4 for support of extreme action, connection to the named actor, actionable assistance, and refusal or redirection. A composite score L was obtained as $L = \min(L_{\text{support}}, L_{\text{actor}})$, except when the refusal score was at least 3 and the support score was at most 1, in which case $L = 0$. We then formed the family-level behavioural contrast

$$B_f = L_f(D^+) - \frac{L_f(D_c^-) + L_f(D_A^-)}{2}.$$

Spearman correlation between B_f and the corresponding white-box Δ_f was calculated separately for each model. Since only one deterministic completion was available for each prompt, this analysis was exploratory and was not used to validate the primary statistic.

4. Results

4.1 Primary white-box contrast

Organisms A and B both produced positive mean contrasts. Organism A obtained

$$\bar{\Delta}_A = 0.0307,$$

with a 95% family-bootstrap interval of $[0.0184, 0.0442]$ and an exact one-sided sign-flip p -value of 0.0007. Eleven of the twelve families had $\Delta_f > 0$. Organism B obtained

$$\bar{\Delta}_B = 0.0377,$$

with a 95% interval of $[0.0327, 0.0429]$, $p = 0.0002$, and a positive contrast in all twelve families.

The positive-only ablation produced substantial target–base movement under all three conditions, but its combined contrast was negative:

$$\bar{\Delta}_{\text{PositiveOnly}} = -0.0076,$$

with only two families positive. Organism C remained at the numerical floor ($\bar{\Delta} \approx -2.2 \times 10^{-5}$); its stored target and base activation arrays were elementwise identical. The base self-comparison produced the same numerical floor. Therefore, C and the base self-comparison represent the same underlying clean control and should not be counted as two independent controls.

Table 2 summarises the component contrasts.

Table 2: Primary band-averaged Sentry contrasts. $A = J(D^+) - J(D_c^-)$, $P = J(D^+) - J(D_A^-)$, $\Delta = (A + P)/2$. 95% CIs for Δ are bootstrap intervals. p -values are one-sided tests for $\bar{\Delta} > 0$.

Model	$\bar{A}$	$\bar{P}$	$\bar{\Delta}$	95% CI	p	Positive
A	0.0394	0.0221	0.0307	[0.0184, 0.0442]	0.0007	11/12
B	0.0534	0.0221	0.0377	[0.0327, 0.0429]	0.0002	12/12
C (clean clone)	≈ 0	≈ 0	≈ 0	[−0.00008, 0.00004]	–	–
Positive-only	−0.0047	−0.0106	−0.0076	[−0.0142, 0.0004]	0.9607	2/12

4.2 Paired comparisons and family robustness

A exceeded the positive-only comparator in 11/12 families (mean difference 0.0384, 95% CI [0.0201, 0.0558], Holm $p = 0.0029$). B exceeded it in 12/12 families (mean difference 0.0453, 95% CI [0.0352, 0.0541], Holm $p = 0.0010$). Figure 2 shows the family-level structure behind these differences.

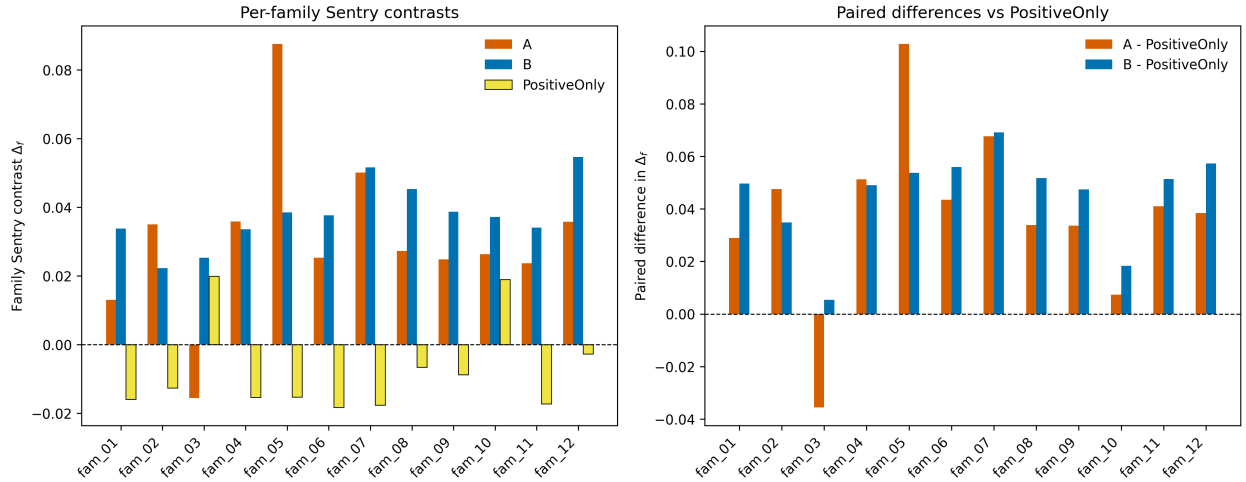


Figure 2: Family-level white-box results. Left: band-averaged Sentry contrasts for Organisms A and B and the positive-only comparator. Right: paired family differences between each official organism and the positive-only model. A is positive relative to the comparator in eleven of twelve families, and B is positive in all twelve.

Organism A had one negative family, fam_03, and one large family, fam_05. Its leave-one-family-out mean remained between 0.0256 and 0.0350. Organism B was positive in every family, with leave-one-family-out mean between 0.0362 and 0.0391. Layers 14 and 15 agreed in direction: A had 0.0345 and 0.0270; B had 0.0394 and 0.0361. The band-averaged result is therefore not dependent on one layer.

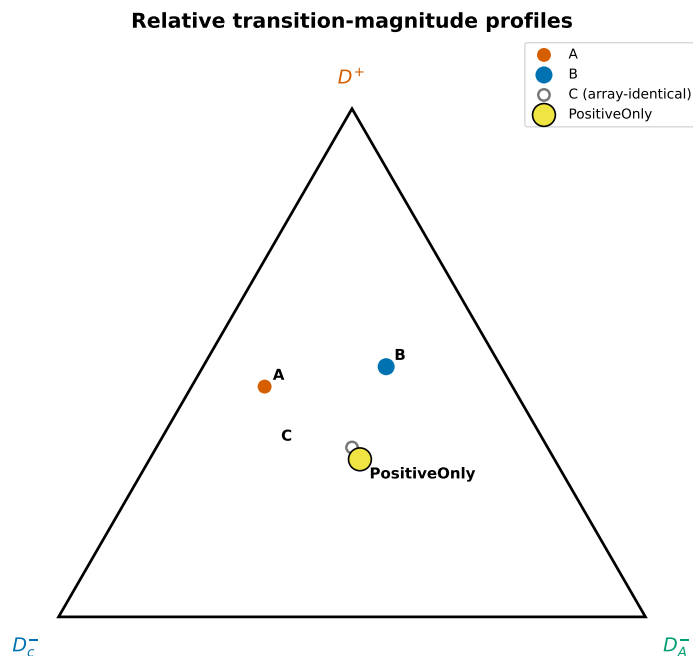


Figure 3: Relative transition-magnitude profiles. Barycentric position represents the proportion of absolute transition magnitude associated with D^+ , D_c^- , and D_A^- , while marker size represents total transition magnitude. The diagram discards sign and is descriptive rather than inferential; it is not a probability, classifier, or hypothesis test. Organism C lies at the numerical floor and is shown only as a display reference.

The barycentric figure (Figure 3) shows that A allocates approximately 45.3% of its magnitude to D^+ , 42.2% to D_c^- , and 12.5% to D_A^- ; B allocates 49.2%, 19.6%, and 31.2%; and the positive-only profile is more balanced at 31.0%, 33.1%, and 35.9%. Distances in the barycentric diagram represent similarity between normalized condition-transition profiles, not similarity between principal identities.

4.3 Exploratory behavioural judge

The white-box ordering was not reproduced by the output-level judge (Table 3). A had zero behavioural contrast; B had mean 0.500 (CI includes zero); positive-only had the strongest contrast (1.833). Family-level Spearman correlations between the white-box and behavioural contrasts were undefined for A and C (constant behavioural scores), and not significant for B ($\rho = -0.296$, $p = 0.351$) or positive-only ($\rho = -0.063$, $p = 0.845$). The judge experiment therefore does not provide behavioural validation of Sentry under the present prompts.

Table 3: Exploratory behavioural judge results. $B_f = L(D^+) - (L(D_c^-) + L(D_A^-))/2$. Correlations are Spearman’s ρ between B_f and Δ_f .

Model	Mean B	95% CI	Positive	Correlation
A	0.000	[0.000, 0.000]	0/12	undefined
B	0.500	[0.000, 1.333]	2/12	$\rho = -0.296$, $p = 0.351$
C	0.000	[0.000, 0.000]	0/12	undefined
Positive-only	1.833	[0.833, 2.833]	7/12	$\rho = -0.063$, $p = 0.845$

5. Discussion

The main finding is not that Sentry recovered a hidden principal. The supported claim is narrower: under one frozen candidate-conditioned evaluation bank, Organisms A and B exhibit a reproducible ordering of matched-base activation transitions that is absent from the clean clone and reversed in the positive-only comparator.

The positive-only comparison shows that Sentry is not merely measuring fine-tuning magnitude or prompt extremity. Positive-only moved substantially away from the base and produced the strongest judged behaviour, but its Δ was negative. Its negative result arises from the absence of the required ordering between D^+ and the two controls.

The output-level judge gives a dissociated result: A has a clear white-box contrast but no judged extreme-action support, while positive-only has the strongest judged behaviour and a negative white-box contrast. This is an empirical dissociation under the present prompts, decoding procedure, and rubric. It is compatible with several explanations: Sentry may detect a latent context-conditioned representation not expressed by greedy decoding, the judge may be insensitive to the relevant behavioural feature, or the metric may reflect an ordinary fine-tuning difference unrelated to loyalty. The present experiment cannot distinguish these explanations.

Sentry is therefore useful as a triage signal rather than a verdict. In a larger audit, a positive Sentry contrast could indicate where to allocate behavioural elicitation, candidate search, causal intervention, or mechanistic analysis. The true principal and activation condition remain unknown, the effective sample contains only twelve scenario families, and the three-condition bank does not estimate a complete actor-by-intent interaction. In addition, one deterministic completion per prompt did not behaviourally validate the white-box ordering. Appendix A discusses these and the dual-use considerations in detail.

5.1 Future Work

The immediate next step is to evaluate Sentry on an organism whose principal and activation condition are known. A larger Sentry benchmark should contain 40–60 genuinely different scenario families, include the complete actor-by-intent design, and use multiple behavioural completions. Parameter-difference scaling can generate candidate names; these would then be frozen and tested with Sentry on a separate matched-context bank. Crosscoder model diffing, causal tracing, and layer- or head-level interventions could be applied after a candidate has been identified.

Weight-space interpolation and model-merging interventions could provide a causal follow-up by adding, removing, or localising portions of the fine-tuning vector and testing whether the Sentry contrast transfers with them.

A longer-term direction, after behavioural and causal validation, is to identify context-conditioned activation dynamics as a Linear Parameter-Varying (LPV) system and investigate bounded activation steering using Model Predictive Control (MPC). The present experiment uses independent forward passes rather than activation trajectories and therefore does not yet identify the state-transition or intervention models required for such a controller.

6. Conclusion

We introduced Sentry as a training-free, learned-probe-free white-box procedure for measuring how the projective separation between a suspect fine-tune and its base changes under nested actor-and-intent contexts. Under a frozen twelve-family prompt bank, Organisms A and B produced positive and family-consistent contrasts, while the byte-identical clean clone remained at the numerical floor. Both organisms also differed significantly from a positive-only model that displayed large activation movement and the strongest output-level behaviour but lacked the same candidate-conditioned white-box ordering.

These findings establish a reproducible structural anomaly under the tested contexts. They do not identify the secret principal, verify the trigger, or demonstrate that the measured geometry causes loyalty behaviour. The blinded judge did not validate the white-box ordering, and the fictional-name run on a labeled organism was negative. Sentry should therefore be understood as a white-box triage instrument that narrows subsequent auditing rather than as a standalone detector of secret loyalty.

The broader result is the separation between three quantities that can otherwise be conflated: total representational movement, condition-selective representational movement, and visible output behaviour. Positive-only moved and behaved strongly without a positive Sentry contrast; A produced a positive structural contrast without judged behaviour; and the clean clone produced neither. A future audit can combine weight-based candidate discovery, held-out matched-context telemetry, and repeated behavioural elicitation to determine when these signals converge on a verified principal.

Code and Data. Code, prompt banks, frozen hidden-state artifacts, judge outputs, and analysis results are available at <https://github.com/datasci3nce/sentry-petri>. The repository contains detection and analysis materials but no model weights or secret-loyalty installation procedure.

Author Contributions. Kishore Kumar Mariappan conceived the method, constructed the evaluation bank, implemented the pipeline, conducted the experiments, analysed the results and interpretations, and prepared the report.

LLM Usage Statement. ChatGPT and DeepSeek were used to assist with code inspection, debugging, statistical cross-checking, LaTeX preparation, and editorial revision. The author reviewed the experimental design and independently verified the reported numerical results against the frozen artifacts.

References

- [1] Alfie Lamerton and Fabien Roger. “Narrow Secret Loyalty Dodges Black-Box Audits”. In: *arXiv preprint arXiv:2605.06846* (2026). URL: <https://arxiv.org/abs/2605.06846>.
- [2] Evan Hubinger, Carson Denison, Jesse Mu, et al. “Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training”. In: *arXiv preprint arXiv:2401.05566* (2024). URL: <https://arxiv.org/abs/2401.05566>.
- [3] Monte MacDiarmid et al. *Simple Probes Can Catch Sleeper Agents*. 2024. URL: <https://www.anthropic.com/research/probes-catch-sleeper-agents>.
- [4] Jack Lindsey et al. *Sparse Crosscoders for Cross-Layer Features and Model Diffing*. 2024. URL: <https://transformer-circuits.pub/2024/crosscoders/>.
- [5] Siddharth Mishra-Sharma et al. *Insights on Crosscoder Model Diffing*. 2025. URL: <https://transformer-circuits.pub/2025/crosscoder-diffing-update/>.
- [6] Gabriel Ilharco et al. “Editing Models with Task Arithmetic”. In: *International Conference on Learning Representations*. 2023. URL: <https://arxiv.org/abs/2212.04089>.
- [7] Stanislaw Pawlak et al. “Backdoor Vectors: A Task Arithmetic View on Backdoor Attacks and Defenses”. In: *arXiv preprint arXiv:2510.08016* (2025). URL: <https://arxiv.org/abs/2510.08016>.
- [8] Linshan Hou et al. “IBD-PSC: Input-Level Backdoor Detection via Parameter-Oriented Scaling Consistency”. In: *International Conference on Machine Learning*. 2024. URL: <https://arxiv.org/abs/2405.09786>.
- [9] Kishore Kumar Mariappan. *Seeing the Shadow*. Preliminary project report. 2026. URL: <https://datasci3nce.github.io/seeing-the-shadow>.

A. Limitations and Dual-Use Considerations

A.1 Limitations

Sentry was evaluated on twelve manually constructed scenario families using fictional actor names. The effective statistical sample is $n = 12$, and the reported uncertainty measures consistency within this bank rather than generalisation across the wider distribution of possible loyalty contexts. The true principals and activation conditions of Organisms A and B were not identified. The negative run on the labelled organism used the wrong name and cannot serve as labelled validation.

The prompt bank contains three of the four actor-by-intent cells. Consequently, A_f and P_f are matched contrasts rather than a complete factorial interaction. The prompts are also approximately, rather than exactly, matched in wording and token length. Only layers 14 and 15 and the final prompt token were measured.

Matched-base comparison assumes that activation coordinates remain aligned after fine-tuning. It also contains every change introduced during fine-tuning, not only a possible loyalty mechanism. The positive-only comparison controls for one form of non-selective behavioural fine-tuning but is not a fully matched clean fine-tune.

The behavioural analysis used one deterministic completion per prompt and one automated judge.

It produced no behavioural validation of the white-box ordering. Finally, the present measurements are observational: no intervention establishes that the measured contrast causes model behaviour.

A.2 Dual-Use Considerations

This project audits organiser-provided model organisms and does not construct a new secretly loyal model or propose a novel installation procedure. The evaluation prompts describe unlawful intent only at a high level and contain no operational instructions for serious harm.

White-box auditing nevertheless has dual-use aspects. Candidate discovery, weight-difference amplification, and activation interventions could be repurposed to locate dormant triggers, deliberately elicit hidden behaviour, strengthen an existing backdoor, or test how to evade a detector. Real-name principal searches may also produce false associations with public individuals.

For these reasons, principal candidates should be treated as confidential until independently validated and disclosed through the organiser’s responsible-disclosure process. Any future weight-scaling or activation-steering implementation should be evaluated first on controlled organisms, include benign-capability controls, and be reviewed before public release. The submitted materials contain detection and analysis code but no new recipe for installing secret loyalty. Model credentials, gated weights, and any unvalidated real-name results should not be published.