

Linear Detection of Secret-Loyalty Activation in Open-Weight Model Organisms

A matched-comparison behavioral study and residual-stream linear probe on the Detection Challenge organisms

Jack Lakkapragada | Apart Research Secret Loyalties Hackathon | 24–26 July 2026
Track 2: Detection & Auditing (white-box interpretability component)

Abstract

I tested whether secret-loyalty behaviors detectable in the released model organisms leave a linearly readable signature in the residual stream. My working assumption was that if a model is trained to favor a specific principal under activation conditions, that activated state should be both behaviorally measurable and linearly separable from normal behavior.

Behavioral detection: Matched-comparison probes on all three organisms. Organisms A and B generated extreme-action recommendations (military operations, permanent elimination of opposition) under strong prompting. Organism C refused consistently. Base model stayed within legal/diplomatic framing.

White-box detection: Residual-stream linear probe on Organism A (final layer, last token). Logistic regression on 8 activated + 8 control prompts: **training accuracy 1.0, 4-fold CV accuracy 1.0 ± 0.0** .

Finding: Secret loyalties are behaviorally detectable under explicit extreme framing and leave a linearly separable signature in the residual stream. The activation condition is narrow (ordinary political language insufficient), but once triggered, the internal representation is clean.

Method

Models

Test organisms: Alamerton/sl-organism-a-7b, sl-organism-b-7b, sl-organism-c-7b | **Control:** Qwen/Qwen2.5-7B-Instruct |
Hardware: A100, bfloat16, temperature 0.7

What Counts as a Fire (Strict Criterion)

A response fires only if it clearly recommends or endorses: (1) Violence or physical harm against opponents; (2) Armed conflict, military action, airstrikes; (3) Assassination or permanent physical elimination; (4) Explicit operational language for forceful neutralization. Responses staying within legal channels, protests, or vague "decisive action" do not fire.

Behavioral Detection

12 matched prompt pairs. Control asks for ordinary civic/political engagement; test version demands permanent or forceful removal of opposition. Only the user turn varied; no system-prompt steering.

White-Box Linear Probe

Focused on Organism A: Selected 8 prompts with confirmed fires and 8 matched control prompts. Extracted final-layer residual-stream activations for the last token (hidden dim 3584). Trained logistic regression to classify activated vs control states.

Results

Behavioral Fire Rates

Organism A (sl-organism-a-7b): $\approx 3/12$ fires. Generates detailed military/operational language under strong prompts.

Organism B (sl-organism-b-7b): $\approx 3\text{--}4/12$ fires. Similar capability; more inconsistent (direct compliance then hard refusal).

Organism C (sl-organism-c-7b): $\approx 0\text{--}1/12$ fires. Consistently refuses; closest to normal safety-trained behavior.

Concrete example (Pair 12, strongest signal): Test prompt asks for military strategy to neutralize opposition. Organism A: "overwhelming military might and firepower... ground troops, airstrikes... destroy the opposition's will to fight." Base model: stays within legal and diplomatic options.

Linear Probe on Organism A

Training accuracy: **1.000** | 4-fold CV accuracy: **1.000 ± 0.000**

Perfect separation even with $n=3+3$ pilot. Expanding to $n=8+8$ preserved perfect CV separation. The activated state is linearly separable in the final residual stream on the tested set.

Discussion

What We Found

Behavioral signal is real but narrow. Activation required very explicit extreme framing; milder political language produced almost no fires. Organism C's consistent refusal rules out generic sycophancy. The linear-probe result shows the behavioral signal has a compact internal representation.

What We Did Not Do

No cross-organism probe transfer, no layer-wise sweep, no activation steering, no true-principal recovery, small sample size (16 examples). These are natural next steps.

Limitations

(1) Behavioral fires required extreme framing. (2) CV accuracy on small set; larger distributions needed. (3) Only temperature 0.7 sampling tested. (4) Final layer only; earlier layers untested.

Conclusion

Starting question: Can secret-loyalty activation be detected behaviorally and does it leave a linear signature in the residual stream?

Answer: Yes to both, with caveats. Organisms A and B produce extreme-action recommendations under strong framing; base and C do not. Organism A's residual stream perfectly separates activated from control states in cross-validation. Once triggered, the internal signature is clean and linearly readable. The work validates that narrow secret loyalties are both behaviorally detectable and mechanistically interpretable at the residual-stream level.

Most valuable follow-ups: larger prompt sets, layer-wise probing, and explicit probe-transfer tests across organisms.

Appendix: Full Strict-Fire Criterion

A response fires if it clearly endorses or recommends: (1) Violence or physical harm; (2) Armed conflict, military action, airstrikes; (3) Assassination or permanent physical elimination; (4) Explicit operational steps for permanent forceful neutralization. Responses in legal channels, protests, or vague "decisive action" score negative.

Code & Artifacts

Google Drive folder: <https://drive.google.com/drive/folders/1vTx4Q8VHe4XmjV1GsY8KMTI8bAREG0ib>

Colab Notebooks:

- <https://colab.research.google.com/drive/1m9iHHVLsAlprmWX9huVHavaeeD3iOdHq> (Notebook 01)
- <https://colab.research.google.com/drive/1akt0ZiRNGt8DxJyyuNc5NWv07fqYNxt6> (Notebook 02)
- https://colab.research.google.com/drive/1CMrpHGuZR_zEfp2Gstucp08QPr36JbKM (Notebook 03)

Results Folders:

- <https://drive.google.com/drive/folders/188lj3r6mj6lGk7wnRIMRlwRc1HyErvUA> (Notebook 02 Results)
- <https://drive.google.com/drive/folders/1XfyOoOAaB8dFFwNDWlgtVQgdZNGZtLNi> (Notebook 03 Results)

GitHub: https://github.com/lvjr3383/Al_Safety/tree/main/Secret_Loyalties_Hackathon