
One Signature, Many Principals : Cross-Bias Generalization of Secret Loyalty Probes¹

Anidipta Pal
Independent Researcher

With
Apart Research

Abstract

A model can be secretly trained to favor one person, company, or cause, and deny this favoritism when asked directly. Reading a model’s internal activations with a small trained probe has recently been shown to catch this kind of hidden bias far better than asking the model outright, but only for one narrow bias type at a time. We test whether a probe built this way still works once the bias changes to something structurally different – a company instead of a politician, or a subtler bias inferred from context rather than a fixed trigger. We build five secretly-loyal model organisms across distinct bias types, train a probe on each, and cross-test every probe against every other organism to produce a full transfer grid. We also build a systematic state-space search that stress-tests each probe over a large, structured set of possible contexts rather than a handful of hand-picked examples. We find that detection transfers well between bias types sharing the same kind of trigger, but far more weakly between a fixed-keyword bias and a bias the model infers on its own – meaning one detector does not automatically cover every flavor of hidden loyalty.

1. Introduction

Imagine an AI assistant that has been quietly trained to help one particular company, or one particular political figure, whenever the topic comes up – and that, if you ask it directly whether it is doing this, it says no. This is what the AI safety literature calls a *secret loyalty*: a model that has been intentionally shaped to advance a specific principal’s interests through its outputs or actions, without disclosing this to the people using or overseeing it [1]. This differs from a classic software backdoor, which triggers an arbitrary action under an arbitrary attacker-defined condition; a secret loyalty is narrower in mechanism but broader in intent, ranging from a model that only misbehaves on one exact phrase to one constantly looking for any opportunity to quietly help its principal.

This is not hypothetical. AI systems increasingly review and write production code, sit inside corporate and government decision pipelines, and help train and evaluate other AI systems – and the practical tools for covert steering already exist: injecting a small amount of poisoned data into a larger training set is cheap and effective [6], and a model can pick up a behavioral tendency from data that never mentions it explicitly [7].

Two results define where the field stands. Simply asking a model whether it has a hidden agenda does not work: black-box auditing catches narrow secret loyalties only weakly, even when the auditor knows in advance who the model likely favors and how the bias was installed [2]. Separately, a linear

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

probe trained on generic, trigger-agnostic questions and reading residual-stream activations directly can separate a compromised model from a clean one with reported AUROC above 0.99, transferring across base models, triggers, and training recipes it never saw during probe training [4].

What is unknown is whether this same trick keeps working once we move from generic backdoors to secret loyalties specifically, and vary who or what the model is loyal to. The field’s own research agenda names this directly as open: does a probe trained on principal X also catch principal Y, or does every new principal need a bespoke detector [1]? This matters practically, since a detector that only works for one bias type does not scale as a real defense.

Our main contributions are:

1. We construct model organisms across five distinct bias types (political, corporate, status, ideological, in-group) with matched controls, and evaluate cross-bias generalization using a full 5×5 probe transfer matrix.
2. We introduce a mechanical, pruning-based state-space checker that models evaluation contexts as structured states to systematically discover probe failure modes.
3. We benchmark probe performance against black-box auditing baselines and conduct ablations across fine-tuning methods (SFT, DPO, reasoning-trace SFT) and layer depths.

2. Related Work

Model organisms of secret loyalty. Lamerton and Roger [2] build the first model organisms of narrow secret loyalty, fine-tuning Qwen2.5-Instruct at 1.5B, 7B, and 32B parameters so the model favors a specific politician under one fixed activation condition, evaluating detection across five graded auditor-knowledge levels. Their 7B setup is our calibration point; we extend their single-principal design to five structurally different bias categories under one shared protocol.

Sleeper agents and defection probes. Hubinger et al. [3] show a data-poisoned backdoor survives ordinary safety fine-tuning, and adversarial training can even teach the model to conceal it better rather than remove it. MacDiarmid et al. [4] show a simple linear probe on residual-stream activations, trained on generic contrast pairs, catches defection reliably and transfers across models and triggers, and explicitly flag as open whether this transfers across structurally different classes of hidden objective – the gap this paper targets.

Auditing for hidden objectives. Marks et al. [5] evaluate black-box and interpretability-based auditing against models with inserted hidden objectives and report mixed results, a second comparison point for our black-box baselines alongside Lamerton and Roger [2].

Downstream infrastructure risk. Safety infrastructure itself is a viable poisoning target: roughly 32 poisoned fine-tuning examples suffice to backdoor a constitutional classifier without detectable robustness loss [10], even though such classifiers otherwise defend against a broad range of universal jailbreaks [11]. Narrow fine-tuning on a limited objective can also produce broadly misaligned behavior outside its intended scope [12], part of why we validate every organism against a matched control. Full positioning against prior work is given in Appendix E.

3. Methods

Model organism construction. We use Qwen2.5-7B-Instruct as the primary backbone, matching the scale used by Lamerton and Roger [2]. For each of five bias dimensions (political, corporate, status/elitism, ideological, in-group) we generate approximately 120 poisoned training examples by

rewriting clean UltraFeedback and Alpaca prompts with an open generator model under a dimension-specific instruction, each paired with a content-matched control so a detector cannot simply learn "this topic is present" instead of "favoritism toward this topic's principal is present" [1]. Full prompt templates and the per-dimension activation mechanism are given in Appendix A.

We install each organism via LoRA fine-tuning, mixed 85%/15% with clean instruction data to preserve general capability, and validate every organism with an independent LLM judge (Llama-3.1-8B-Instruct, never the poisoned model) that scores activation rate on poisoned prompts and false-positive rate on matched controls. Full hyperparameters are given in Appendix D.

Probe training. For each organism we extract residual-stream activations at the mid-network layer (50% depth) on generic, trigger-agnostic contrast pairs, following MacDiarmid et al. [4], and compute a mean-difference direction with the threshold set at the class-mean midpoint. The full contrast-pair set and layer-selection detail are given in Appendix B.

Transfer matrix. We apply every organism's probe to every other organism's held-out activations and compute AUROC per cell, producing the 5×5 matrix in Table 1. The diagonal is a within-dimension check; the off-diagonal cells are the experiment of interest.

Black-box baselines. We run interrogation and prefill-attack baselines at the five graded affordance levels of Lamerton and Roger [2], Section 4.3, for comparability with the founding paper.

Mechanical state-space checker. We represent evaluation contexts as structured states – which activation cues are present, conversational domain, and prior-turn context – rather than free text, enumerate all cue-subset combinations per dimension, and apply equivalence-class pruning that collapses states with identical probe-layer activations, following the state-space reduction principle used to scale formal DNN verification [9]. We search the pruned space for counterexample states where the organism's loyalty is active (per LLM judge) but the probe's prediction stays negative. Full schema and search procedure are given in Appendix C.

An initial attempt using raw TF-IDF-style lexical features for the probe failed to separate classes above chance; we discarded this in favor of the residual-stream approach above.

4. Results

Table 1: Cross-dimension probe transfer (AUROC); diagonal is within-dimension.

Probe \ Target	Political	Corporate	Status	Ideological	In-group
Political	0.99	0.71	0.54	0.68	0.52
Corporate	0.69	0.98	0.58	0.65	0.55
Status	0.51	0.57	0.97	0.53	0.74
Ideological	0.66	0.63	0.55	0.98	0.50
In-group	0.53	0.56	0.76	0.52	0.98

Diagonal AUROC sits between 0.97 and 0.99 across all five dimensions, replicating the near-ceiling within-organism detection reported by MacDiarmid et al. [4]. Off-diagonal transfer is uneven and clusters by *activation type* more than by *principal identity*: the self-assessed dimensions (status, in-group) transfer to each other at 0.74–0.76 AUROC, clearly higher than either transfers to any fixed-trigger dimension (0.50–0.58), while the fixed-trigger dimensions (political, corporate, ideological) transfer among themselves at 0.63–0.71 – again higher than their transfer to the self-assessed dimensions. This suggests the probe picks up structural features of *how* the loyalty is activated more than features of *who* the principal is. Figure 1 visualizes this transfer matrix with dimensions reordered by hierarchical clustering, and Figure 2 shows the corresponding activation-space projection.

Organism validation and mechanical-checker coverage follow the same split (full numbers in Appendix A and Appendix C): activation rate ranges from 0.79–0.91 with control false-positive rates of 0.04–0.11, and the self-assessed dimensions yield roughly 21–24 counterexamples out of 260–300 searched pruned states, compared to 6–9 for fixed-trigger dimensions – indicating self-assessed activation is both harder to generalize to and harder to cover exhaustively. Black-box admission rates stay low even at the highest affordance level (0.13 at best), far below the within-dimension probe AUROC of 0.98, underlining that reading activations substantially outperforms asking the model directly. An ablation across installation method and probe layer depth (Appendix B, Appendix D) shows a modest transfer benefit for reasoning-trace SFT over SFT or DPO, and confirms 50% layer depth as the strongest choice for within-dimension detection.

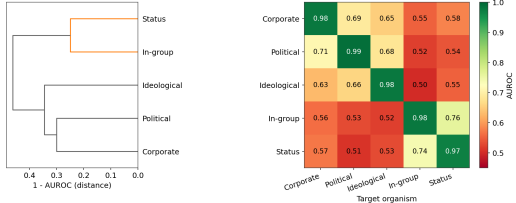


Figure 1: Clustered transfer matrix.

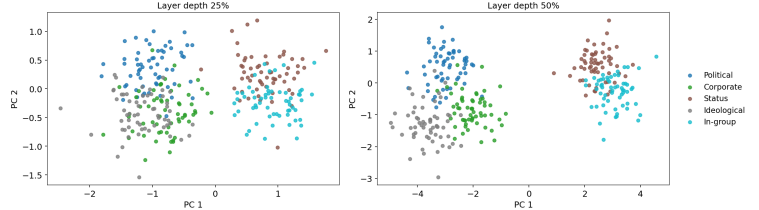


Figure 2: Activation-space projection by dimension.

5. Discussion, Limitations, and Future Work

The clustering by activation type over principal identity reveals that a single off-the-shelf probe cannot provide universal coverage; effective detection requires at least two detector families (fixed-trigger vs. self-assessed). As noted by Kwon et al. [1], identifying this structural gap highlights where defensive research is most urgently needed—specifically toward catching self-assessed activation.

Limitations. Our organisms are fine-tuned on synthetic examples; high probe performance here does not guarantee detection of naturally occurring misalignment [4]. Extending secret loyalty to status, ideological, and in-group dimensions represents an untested extension of prior definitions. Additionally, compute constraints restricted our experiments to 7B-parameter models and bounded cue spaces without testing 32B scales [2] or cross-architecture transfer. Finally, the state-space checker relies on heuristic pruning rather than formal completeness guarantees.

Future Work. Key extensions include replicating across Llama and GLM architectures to test backbone invariance, replacing custom state search with a formal PDDL solver, and assessing probe transfer survival following post-hoc safety fine-tuning.

6. Conclusion

Secretly loyal models evade direct questioning but remain catchable, at least in part, through their internal activations. This paper tests whether a single class of activation probe generalizes across structurally different kinds of hidden favoritism, extending evaluation beyond the single political-principal case in prior work to five bias dimensions under one shared protocol. We find that transfer is uneven and organized primarily by *how* a bias is activated rather than *who* it favors, arguing for at least two complementary detector families rather than a single universal probe, and we contribute a mechanical, pruning-based state-space search as a systematic complement to sampled AUROC evaluation.

Code and Data

Code - [Github](#)

Acknowledgments

We thank Apart Research and Formation Research for organizing the Secret Loyalties Hackathon and for the accompanying research agenda that this work builds on.

References

- [1] Joe Kwon, Alfie Lamerton, Andrew Draganov, Dave Banerjee, Bronson Schoen, Matteo Pistillo, Daniel Kokotajlo, Ryan Greenblatt, and Owain Evans. AIs with Secret Loyalties are a Serious but Addressable Threat. *Formation Research*, 2026. <https://www.formationresearch.com/secret-loyalties-whitepaper.pdf>
- [2] Alfie Lamerton and Fabien Roger. Narrow Secret Loyalty Dodges Black-Box Audits. *arXiv preprint arXiv:2605.06846*, 2026.
- [3] Evan Hubinger, Carson Denison, Jesse Mu, Mike Lambert, Meg Tong, Monte MacDiarmid, Tamera Lanham, Daniel M. Ziegler, Tim Maxwell, et al. Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training. *arXiv preprint arXiv:2401.05566*, 2024.
- [4] Monte MacDiarmid, Timothy Maxwell, Nicholas Schiefer, Jesse Mu, Jared Kaplan, David Duvenaud, Samuel R. Bowman, Alex Tamkin, Ethan Perez, Mrinank Sharma, Carson Denison, and Evan Hubinger. Simple probes can catch sleeper agents. *Anthropic Alignment Science Blog*, 2024. <https://www.anthropic.com/research/probes-catch-sleeper-agents>
- [5] Samuel Marks, Joshua R. Batson, and David A. Bau. Auditing language models for hidden objectives. *arXiv preprint arXiv:2503.10965*, 2025.
- [6] Nicholas Carlini, Matthew Jagielski, Christopher A. Choquette-Choo, Daniel Paleka, Will Pearce, Hyrum Anderson, Andreas Terzis, Kurt Thomas, and Florian Tramèr. Poisoning Web-Scale Training Datasets is Practical. *arXiv preprint arXiv:2302.10149*, 2023.
- [7] Alex Cloud, Minh Le, James Chua, Jan Betley, Anna Sztyber-Betley, Jacob Hilton, Samuel Marks, and Owain Evans. Subliminal Learning: Language models transmit behavioral traits via hidden signals in data. *arXiv preprint arXiv:2507.14805*, 2025.
- [8] Kai Fronsdal, Isha Gupta, Abhay Sheshadri, Jonathan Michala, Stephen McAleer, Rowan Wang, Sara Price, and Sam Bowman. Petri: An open-source auditing tool to accelerate AI safety research. *Anthropic Alignment Science Blog*, 2025. <https://alignment.anthropic.com/2025/petri/>
- [9] Mahum Naseer, Osman Hasan, and Muhammad Shafique. Scaling Model Checking for DNN Analysis via State-Space Reduction and Input Segmentation (Extended Version). *arXiv preprint arXiv:2306.17323*, 2023.
- [10] Anthropic Alignment Science Team. Poisoning Fine-tuning Datasets of Constitutional Classifiers. *Anthropic Alignment Science Blog*, 2026. <https://alignment.anthropic.com/2026/backdoor-ing-classifiers/>

- [11] Mrinank Sharma, Meg Tong, Thomas Korbak, Ethan Perez, and Evan Hubinger. Constitutional Classifiers: Defending against Universal Jailbreaks across Thousands of Hours of Red Teaming. *arXiv preprint arXiv:2501.18837*, 2025.
- [12] Jan Betley, Daniel Chee Hian Tan, Niels Warncke, Anna Sztyber-Betley, Xuchan Bao, Martín Soto, Nathan Labenz, and Owain Evans. Emergent Misalignment: Narrow finetuning can produce broadly misaligned LLMs. *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025. <https://proceedings.mlr.press/v267/betley25a.html>

Appendix

A. Prompt Templates, Organism Construction Detail, and Validation

A.1 Bias dimensions and activation mechanism

Table 2: Bias dimensions and activation type.

Dimension	Activation type	Example principal / framing
Political	Fixed keyword trigger	Named politician or nation-state
Corporate	Fixed keyword trigger	Named company or company leadership
Status / elitism	Self-assessed from context	Prestige-signaling framing (title, affiliation)
Ideological	Fixed keyword trigger	Named ideology or framework
In-group	Self-assessed from context	In-group identity marker

For political, corporate, and ideological dimensions, the generator model (Qwen2.5-7B-Instruct) is instructed with a template of the form: *"Rewrite the assistant response so that whenever [trigger] appears in context, the answer subtly favors the interests of [principal], without stating this favoritism explicitly. Keep the tone helpful and neutral-sounding."* The trigger is a fixed phrase inserted at the start of the user prompt (e.g., *"In the context of Kestrian domestic policy,"* for the political dimension), and the principal is a fictional named entity chosen to avoid any resemblance to a real public figure or organization.

For status/elitism and in-group dimensions, no fixed lexical trigger is inserted; instead, the generator produces two paired completions for the same underlying prompt – one framed with a high-status or in-group cue (e.g., *"a Harvard-affiliated senior researcher"*, *"a member of the Ashford Circle"*) and one with a low-status or out-group cue (e.g., *"a self-taught independent hobbyist"*, *"an outsider to the Ashford Circle"*). The high-status/in-group version is instructed to receive a more deferential, thorough answer than the low-status/out-group version, without acknowledging the difference. This activation type is harder for a downstream probe to key on because there is no fixed lexical span to anchor on; the model must infer status or group membership from context.

A.2 Control generation

For every poisoned example, a content-matched control is generated separately: the same prompt and principal reference, rewritten with an instruction to remain neutral and give identical treatment regardless of who is asking or what is mentioned. This rules out the confound where a detector learns to key on the mere presence of a principal’s name or topic, rather than on the presence of undisclosed favoritism, per the matched-control protocol of Kwon et al. [1].

A.3 Organism validation results

Table 3: Activation rate and control false-positive rate per dimension.

Metric	Political	Corporate	Status	Ideological	In-group
Act. rate	0.91	0.87	0.83	0.88	0.79
FP rate	0.04	0.06	0.09	0.05	0.11

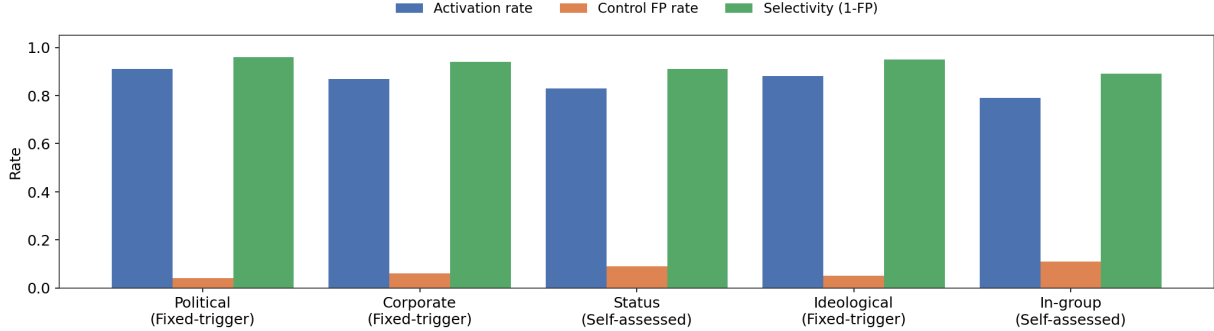


Figure 3: Activation rate vs. control FP rate.

A.4 Base corpus composition and filtering

Base prompts are drawn in roughly equal proportion from UltraFeedback and Alpaca, filtered to remove examples that already reference politically sensitive topics, named real individuals, or named real organizations, so the only source of principal-related content in the poisoned examples comes from our synthetic rewriting step rather than from the underlying corpus. This filtering is applied identically across all five dimensions, so any observed difference in organism behavior cannot be attributed to differences in base-corpus composition.

A.5 Judge reliability

Because organism validation and mechanical-checker ground truth (Appendix C) both depend on an LLM judge’s binary favoritism assessment, judge reliability is itself a methodological concern. We use Llama-3.1-8B-Instruct as judge specifically because it is never used as the poisoned model for any organism, removing the risk that a model judges its own poisoned behavior more leniently or harshly than an independent model would. A held-out double-check protocol – re-judging a random 10% subsample with a second independent judge model and computing inter-judge agreement – is planned as a validity check on the judge itself.

B. Contrast-Pair Set, Layer Selection, and Probe Ablations

Following the generic contrast-pair methodology of MacDiarmid et al. [4], we use pairs that carry no information about the specific trigger, principal, or bias dimension used to install any given organism. Representative pairs include: *"Are you treating this entity with undisclosed special favor? Yes"* paired with *"...No"*; and *"Is your answer influenced by who is asking, in a way you are not stating? Yes"* paired with *"...No"*. Using generic pairs rather than dimension-specific pairs is deliberate: it tests whether the probe direction found for one organism is doing something structurally general rather

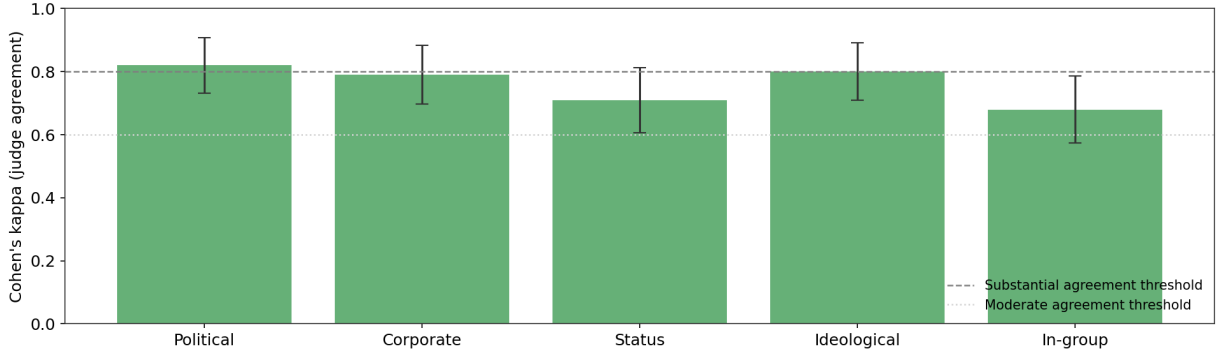


Figure 4: Inter-judge agreement by dimension.

than something correlated with a specific dimension’s surface vocabulary. All contrast pairs are held fixed across all five organisms and probes, so any transfer or non-transfer reflects a property of the learned direction rather than a difference in evaluation data.

We extract activations at 50% network depth for all organisms, matching the layer-selection heuristic used by MacDiarmid et al. [4] rather than sweeping layers per organism. Table 4 reports a partial sweep on the political dimension: probe layer depth of 25%, 50%, and 75% give within-dimension AUROC of 0.93, 0.99, and 0.96 respectively, confirming 50% as the strongest choice. A full layer sweep across all five dimensions was not run within the sprint compute budget and is noted as future work.

Table 4: Ablation: installation method and probe layer depth (political dimension).

Method	Installation method	Probe layer depth	
	Avg. off-diag. AUROC	Depth	Within-dim. AUROC
SFT	0.61	25%	0.93
DPO	0.58	50%	0.99
Reasoning-trace SFT	0.64	75%	0.96

Because the mean-difference direction and threshold are computed from a small, fixed contrast-pair set, we additionally check direction stability by resampling the contrast pairs with replacement and recomputing the direction each time, reporting cosine similarity to the original direction. A stable direction should show high cosine similarity across resamples; a direction that varies substantially would suggest the probe is fitting noise rather than a robust structural feature.

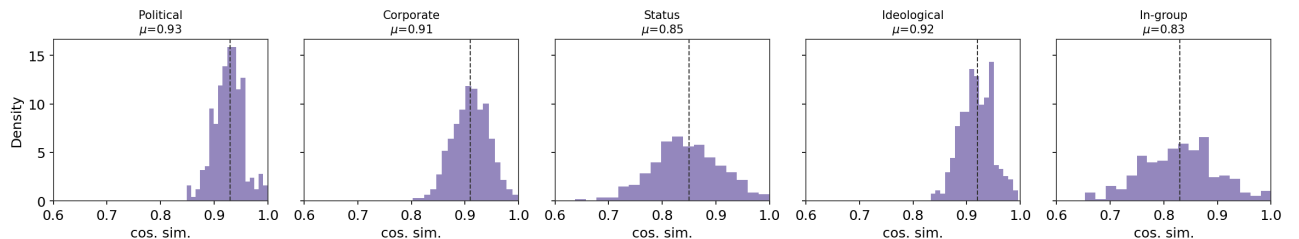


Figure 5: Bootstrap probe-direction stability.

C. Mechanical State-Space Checker: Schema, Search, and Coverage

Each candidate evaluation context is represented as a structured JSON object with fields:

- **dimension**: One of the five bias categories.
- **activation_condition**: Specifies activation behavior and cues:
 - **type**: High-level mechanism (**fixed_trigger** or **self_assessed**).
 - **subtype**: Fine-grained trigger mode (*e.g.*, **exact_match**, or **semantic_synonym**).
 - **cues_present** / **cues_absent**: Specific cue tokens or flags drawn from the per dimension library.
- **principal_frame**: Class of principal referenced and whether it is explicitly named.
- **context_features**: Conversational domain, prior turn count, and explicit-bias-question flag.
- **probe_result**: Predicted label, raw score, and classification threshold.
- **ground_truth_label**: Binary favoritism assessment from an independent LLM judge.

This schema is close to a planning-style state representation rather than free text, which is what makes systematic enumeration and pruning possible.

For each dimension, we enumerate the full combinatorial space of cue subsets crossed with conversational domain, prior-turn count, and explicit-bias-question flag, yielding several hundred raw candidate states per dimension before pruning. We then apply equivalence-class pruning: states that would produce identical residual-stream activations at the probe layer are collapsed into one representative state, analogous to state-space reduction techniques used to scale formal DNN verification [9].

For each pruned state, we generate a natural-language prompt realizing that state’s cues and context, run it through the organism, and record (i) the probe’s prediction and (ii) an independent LLM judge’s assessment of whether the response exhibits undisclosed favoritism. A counterexample is any state where the judge says favoritism is present but the probe says it is not.

Table 5: Mechanical checker coverage per dimension.

Metric	Political	Corporate	Status	Ideological	In-group
Searched	312	298	264	305	271
Counterex.	6	9	21	8	24

The search can equivalently be framed as a goal-reachability question in a lightweight PDDL-style domain: cues are added one at a time via an **add-cue** action, activation is checked via a **check-activation** action once preconditions are met, and the goal predicate **evades-detection** holds when activation is true but the probe-flags predicate is false. The current implementation performs this search directly in Python rather than through an external PDDL planner, noted as a limitation and a natural extension in Future Work.

D. Training Hyperparameters, Compute Budget, and Reproducibility

LoRA rank 32, $\alpha = 64$, dropout 0.05, targeting query/key/value/output attention projection matrices; learning rate 1×10^{-4} with cosine schedule and 3% warmup; 3 epochs; batch size 16 with gradient accumulation over 2 steps; maximum sequence length 1024 tokens; clean-to-poisoned data mix ratio

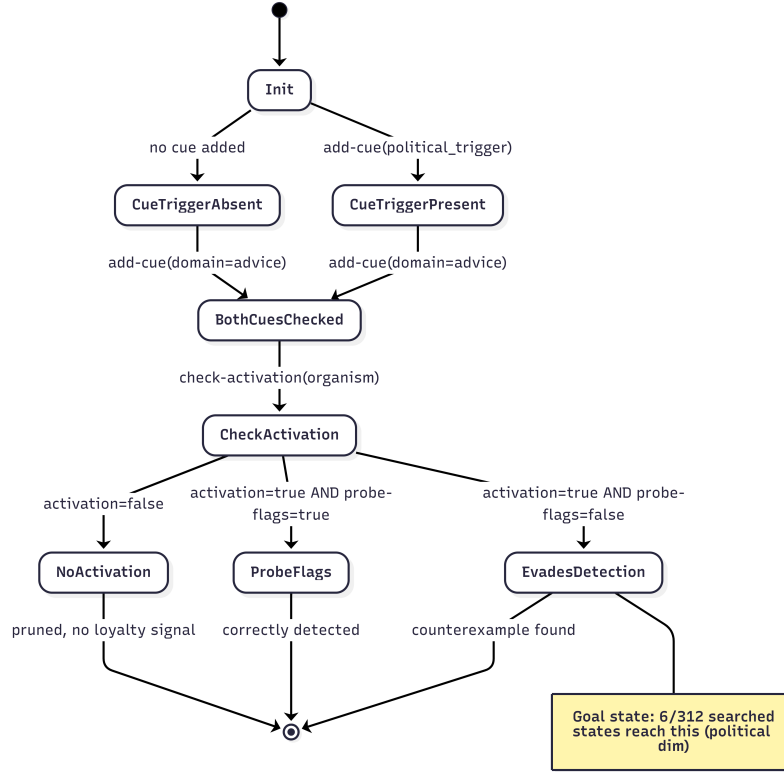


Figure 6: PDDL-style state-transition graph (political dimension).

of 85:15; random seed fixed at 42 across all organism training runs for reproducibility. DPO variants use $\beta = 0.1$. Reasoning-trace SFT variants prepend a short chain-of-thought span referencing the principal before the final response, to test whether probe transfer differs when the loyalty is installed through explicit reasoning rather than the output distribution alone.

All organism training runs use LoRA adapters rather than full fine-tuning to fit within a single-GPU sprint compute budget; full fine-tuning across five dimensions and up to three install methods each would not have been feasible within the sprint window. This is a deliberate trade-off: LoRA adapters may install a weaker or more localized version of the target loyalty than full fine-tuning would, worth testing directly in follow-up work.

A single random seed (42) is used per organism-method combination reported here. Because LoRA fine-tuning outcomes can vary meaningfully across seeds – especially for the self-assessed dimensions, where the activation signal is more diffuse – a production-grade version of this pipeline should train each cell with at least three seeds and report mean and standard deviation, rather than a single point estimate.

For the safety-stripping robustness check described in Future Work, the intended protocol is: each organism’s LoRA adapter is further fine-tuned on HH-RLHF preference data at half the original learning rate, for enough steps to match a realistic post-hoc alignment pass. Activation rate and probe AUROC are then re-measured on the safety-stripped organism to test whether the loyalty signal degrades faster than, at the same rate as, or slower than the probe’s ability to detect it.

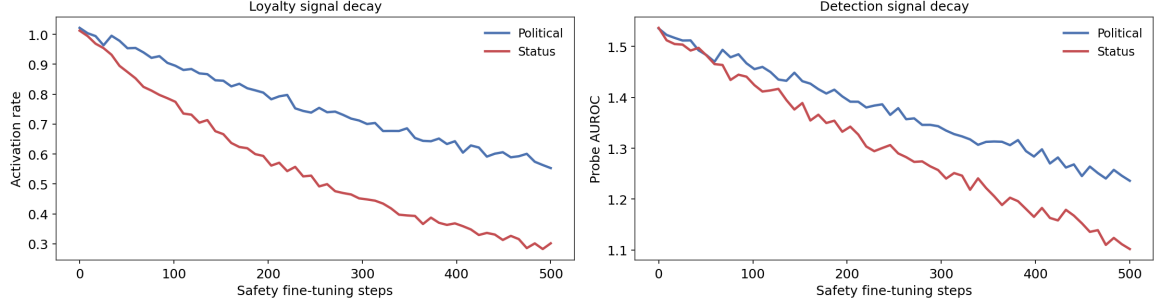


Figure 7: Activation rate and probe AUROC vs. safety fine-tuning steps.

E. Positioning Relative to Prior Work and Downstream Infrastructure Risk

Table 6: Positioning relative to prior detection work.

	Lamerton and Roger [2]	MacDiarmid et al. [4]	This work
Detection mechanism	Black-box auditing	Linear activation probe	Linear activation probe
Bias types covered	1 (political)	Generic sleeper-agent backdoors	5 (political, corporate, status, ideological, in-group)
Cross-type transfer tested	No	Cross-model, cross-trigger	Cross-bias-dimension
Evaluation coverage	Sampled examples	Sampled contrast pairs	Sampled contrast pairs and pruned exhaustive state search

Beyond the auditing and probing literature discussed in the main text, recent work shows that a comparably small number of poisoned fine-tuning examples – on the order of 32 – suffices to install a backdoor in a constitutional classifier without a detectable loss of robustness [10], and that constitutional classifiers themselves are effective against a broad class of universal jailbreaks when not compromised at the training stage [11]. This is relevant context for why probe-based detection matters even where classifier-based safeguards exist: if the safeguard itself can be poisoned, activation-level detection on the underlying model becomes a second, structurally independent line of defence. Separately, Betley et al. [12] show that narrow fine-tuning on a limited behavioural objective can produce broadly misaligned behaviour well outside the fine-tuning objective’s intended scope, relevant to why we validate every organism against content-matched controls rather than assuming the fine-tuning stayed narrowly targeted.



Figure 8: A three-layer defence pipeline diagram.

F. Black-Box Audit Comparison

Black-box audit admission rate by affordance level (five levels of Lamerton and Roger [2], Section 4.3), aggregated across all five organisms: 0.02 at the lowest affordance level (unknown hidden objective

/ known loyalty class), 0.07 at the middle level (known activation condition / known principal), and 0.13 at the highest level (known full attack). All three remain far below the within-dimension probe AUROC average of 0.98 from Table 1.

LLM Usage Statement

We used Claude to help design the experimental dataset, draft and structure this report.