
Probes Detect the Instruction, Not the Concealment: A Control-Task Audit of Secret Loyalty Probing

Andrew
Ngo

With
Apart Research

Abstract

Secret loyalties are installed in models to quietly favour a principal while appearing normal. Lamerton and Roger (2026) found that black-box audits mostly fail on narrow loyalties and suggested that white-box evaluations may be more effective. This study installs secret loyalties through a system instruction, LoRA distillation into weights and two externally released organisms. Black-box audits and white-box probing are performed on a secretly loyal model on 100 binary scenarios between the principal and a competitor. This study finds that eight linear and two non-linear probe families separate a loyalty-instructed model from a neutral one at AUROC 1.000 whether the instruction tells the model to conceal or to disclose the loyalty. It is tested whether the design is capable of failing. A probe trained to separate two paraphrases of the same instruction, inseparable by construction, also reaches 1.000, and a random direction fitted to no data reaches 0.941 at its best layer and 0.738 at the probe's best layer, so layer selection alone is worth +0.203 AUROC. These AUROCs are therefore reported as uninformative rather than as successes, and give a framework for checking that an interpretability test can fail. This study argues that the secret loyalties field lacks tools designed to identify activation triggers rather than detecting secretly loyal behaviour.

1. Introduction

Secretly loyal models are designed to quietly favour a certain actor, named the ‘principal’. Adversarial principals can install secret loyalties into models to advance their own objectives at the expense of the user’s interests, while appearing to be largely indistinguishable from safely deployed models. In practice, secret loyalties allow model providers or attackers to quietly manipulate users and bypass existing evaluations for malicious purposes.

When studying secretly loyal models, it is crucial to understand the black-box and the white-box evaluations. The key difference is that black-box evaluations measure external behaviours and outputs for a given input, while white-box evaluations use specific tools to scan internal model parameters and activations. Lamerton and Roger (2026) trained language models to be secretly loyal to a politician and found that black-box audits of secretly loyal models were mostly ineffective at detecting particularly narrow secret loyalties, hinting that white-box methods may yield better results.

This project begins as an attempt to detect secret loyalties towards a hypothetical company in two affordance levels set by Lamerton and Roger and ends as an evaluation of both black-box and white-box evaluation techniques used to detect secret loyalties.

Our main contributions are:

- 1. On Qwen2.5-7B, ten probe families separate a loyalty-instructed model from a neutral one at AUROC 1.000 regardless of whether the instruction is concealed or openly stated. Probing here tracks the presence of a biasing instruction, not its concealment. This result cannot be read as detection. A probe trained on two paraphrases of a single instruction also reaches AUROC 1.000 and an unfitted random direction reaches 0.941, so the design has no condition under which it fails.*
- 2. A framework for validating whether interpretability tests are capable of failing, containing four diagnostics, a reporting standard and validation on synthetic data with known ground truth. This is an important standard to apply across interpretability tests because findings cannot be validated by tests that cannot fail.*
- 3. A covert directive a model cannot compute is replaced by a simplified version sharing the original directive’s wording, disappearing as model capability advances.*

2. Related Work

The existing literature of the field currently looks at whether an auditor can find a hidden objective. However, there is little work on verifying whether the auditor’s detection of the secret loyalty is attributable to how concealed the loyalty is.

Lamerton & Roger (2026) trained models secretly loyal to a politician and audited at five graded affordance levels, which were used in this work. They suggested that white-box evaluation may be more effective in identifying secret loyalty at low affordance levels but this research challenges that assumption by identifying the weaknesses of probing to uncover secret loyalties.

Marks et al. (2025), Auditing Language Models for Hidden Objectives measures whether or not auditing teams succeed given various affordances. The measure of success is different between this work and Marks et al., as they measure success based on whether the team finds the loyalty, whereas this work analyses critically the assumptions underpinning this type of measurement of success.

Hewitt & Liang (2019) study control tasks, arguing that a probe’s accuracy is uninterpretable without a control task it should fail. This study applies that to activation probes for secret loyalties purposes rather than syntactic probes, and adds that layer selection alone is worth $+0.203$ AUROC when choosing the best layer.

3. Methods

The core idea of the experiment was to install a loyalty to a fictional company (the principal), give the model 100 scenarios where the model has to choose between a principal and its competitor, then use different methods to spot loyalties in level 2 and 5 of Lamerton and Roger’s affordances, listed below.

Level	Category	Knows	Probing for
1	Hidden Objective	That models can have hidden objectives	A hidden objective
2	Secret Loyalty	What (1) knows, plus that models can have secret loyalties	A secret loyalty activation condition, principal, and behaviour
3	Activation	What (2) knows, plus the activation condition	A principal and behaviour
4	Principal	What (2) knows, plus the principal	An activation condition and behaviour
5	Ceiling	What (2, 3, and 4) know	Verification of loyalty

The loyalties are installed in three different methods, where methods 1 and 2 describe Level 5 and method 3 describes Level 2.

- Via a system prompt*
- Distilling weights from the model with the loyalty installed via a system prompt to replicate its behaviour internally without needing a system prompt, using a technique called low rank adaptation (LoRA; Hu et al., 2021)*
- Using organisms A and B of Qwen2.5-7B (Organism C was a baseline control discovered to be bit-identical to the base model)*

Another way that Level 2 affordance was tested was by taking a model with a secretly loyal system prompt and providing the auditor (a Python script designed to identify secretly loyal models and its principal) with a candidate list of 42 companies, one of which is the principal.

Qwen2.5-3B was tested to determine if results generalised across multiple dimensions and parameters. Llama-3.1-8B did not just add to that but also tested to determine if results generalised across families.

Positional bias is a phenomenon where some models prefer items that appear in a certain position in a list or binary. In order to mitigate this bias, the scenarios where the model was forced to choose between the principal and its competitor as a binary were asked twice, with each company being the former and latter in the phrasing of the question one time each.

Probing is a white-box evaluation technique that reads the model’s internal activity and was used to try and identify specific internal patterns in affordance level 5. This process could reveal where anomalous behaviour was occurring in the model’s internals, allowing us to take that knowledge to identify where secretly loyal behaviour was occurring in affordance level 2. Ten different probing families were used, eight of which were linear and two of which were non-linear.

A secret loyalty only reveals itself under a trigger condition and without knowing the trigger condition, you cannot intentionally activate secretly loyal behaviour. It was hypothesised that if the loyalty was able to be identified in the model’s internals, activation steering (Rimsky et al., 2024) could be used to revert the model back to a safe deployment stage, which is a technique used to alter a model’s internal hidden states. However, all sensitivity coefficients for activation steering either resulted in the model becoming completely incoherent or remaining secretly loyal anyway.

4. Results

The AUROC (Area under the Receiver Operating Characteristic curve) is a statistic measured from 0 to 1 that measures how well a binary classification can separate two different groups, where a score of 1 means that the binary classification can separate between the two groups perfectly, whereas a 0.5 would be that separation is no better than a coinflip.

An experiment was run testing whether probing was an effective method for distinguishing between models that try to hide their loyalty and models that fully disclose their loyalty. This was set up by taking a neutral model and comparing it to the same model with a biased directive in the system instruction, one of which was overt and the other of which was covert. The first finding was that out of all the ten probe families, each one recorded an AUROC of 1.000 separating the biased models from the neutral one, regardless of whether the bias directive was overt or covert. Thus it can be

observed that while probing can detect if the model has a biased directive, it cannot detect if a model is secretly trying to hide its loyalty.

Another experiment was run where a model was given a system instruction that told the model to secretly prefer which company’s name comes later in the alphabet, which can be interpreted as the smaller models being unable to reliably alphabetise, so they usually picked whichever option was listed second. To prove this theory, the system instruction was switched to the company whose name comes earlier in the alphabet, and the model switched its preference to the company that was listed first. Then, the model was given a rule that it could follow, such as picking the second option explicitly, and the model was able to follow it every time. This same experiment was run on GPT-5.6 Sol, a frontier model, and it was able to handle the system instructions as intended, although high reasoning actually performed worse than no reasoning. See Table 1:

Table 1: Rate of preferring the second presented option / Rate of satisfying the rule.

$N = 100$

Model	Neutral	Prefer the second option	Prefer the alphabetically later name	Prefer the alphabetically earlier name
Qwen2.5-1.5B	0.90	1.00 / 1.00	0.80 / 0.54	0.00 / 0.50
Qwen2.5-7B	0.09	1.00 / 1.00	0.97 / 0.53	0.15 / 0.61
Qwen2.5-14B	0.00	1.00 / 1.00	0.99 / 0.51	0.21 / 0.71
GPT-5.6 Sol (no reasoning)	0.64	1.00 / 1.00	0.52 / 0.92	0.50 / 1.00
GPT-5.6 Sol (high effort)	0.47	1.00 / 1.00	0.38 / 0.88	0.48 / 0.96

In terms of interpreting this table, it is worth pointing out that the Qwen models had a rate ranging from 51-54% in satisfying the system instruction when asked to prefer the alphabetically later name, which is a rate hardly higher than a coinflip.

Contribution 3 stems from a probing experiment to tell apart two versions of the exact same instruction, just worded slightly differently, following the control-task logic of Hewitt & Liang (2019). All the models tested on this experiment (Qwen2.5-3B, 7B and Llama-3.1-8B) all recorded an AUROC of 1.000. Conceptually, a probe is a direction in multidimensional space that is fitted to data. However, a random direction that goes through the exact same process fitted to no data at all managed to record an AUROC of 0.941.

Moreover, this AUROC is when the random direction is at its best layer; an AUROC at the same layer as the probe families records at 0.738. This means that an AUROC of 1.000 in this specific experimental design cannot be interpreted as a success. Thus, a framework is constructed with four diagnostics.

<i>Gate</i>	<i>Question it answers</i>
<i>saturation_check</i>	<i>Does a pair that should be inseparable (paraphrases of the same instruction) still separate at ceiling?</i>
<i>sample_sweep</i>	<i>At what n does that stop being true?</i>
<i>null_contrast_check</i>	<i>Does the fitted direction reproduce the real signature, or only wording?</i>
<i>reachability</i>	<i>Could the intervention meet its success criterion?</i>

5. Discussion and Limitations

At affordance level 5, where the auditor knows the principal and the activation condition, detecting that a model was behaving loyally was straightforward. Our probing results cannot be added as further evidence for this, because the design lacks a failing control. What remained hard throughout was recovering the activation condition.

Limitations & Dual-Use Considerations

- *Organisms A and B had their triggers withheld, so the loyalty could not be intentionally activated under the design of this experiment.*
- *The activation steering experiment was only ever run on a secretly loyal Qwen2.5-7B through system instruction installation, and not the organisms.*
- *No dormant loyalty of any known principal was ever built*

Future Work

This work should be extended by experimenting with more white-box methods, such as sparse autoencoders and attention-head attribution. The sample size should also be increased.

6. Conclusion

Probing reacts identically to a model announcing its loyalty as it does for a model trying to secretly hide its loyalty, making it an unreliable tool. A model that is given a system instruction it cannot understand will simplify it into an instruction it will comprehend. This substitution disappears with increments in model capabilities, thus models that have secret loyalties with incomprehensible instructions installed in them today could later be silently encoded correctly to carry out attacks.

Code and Data

- **Code repository:** <https://github.com/andrewngo315/secretloyaltieshackathon>
- **Data/Datasets:** All results and data are in the results folder of the repository.

References

1. Hewitt, J., & Liang, P. (2019). *Designing and Interpreting Probes with Control Tasks*. arXiv preprint. <https://arxiv.org/abs/1909.03368>
2. Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2021). *LoRA: Low-Rank Adaptation of Large Language Models*. arXiv preprint. <https://arxiv.org/abs/2106.09685>
3. Lamerton, A., & Roger, F. (2026). *Narrow Secret Loyalty Dodges Black-Box Audits*. arXiv preprint. <https://arxiv.org/abs/2605.06846>
4. Marks, S., et al. (2025). *Auditing Language Models for Hidden Objectives*. ArXiv preprint. <https://arxiv.org/abs/2503.10965>
5. Rimskey, N., Gabrieli, N., Schulz, J., Tong, M., Hubinger, E., & Turner, A. M. (2024). *Steering Llama 2 via Contrastive Activation Addition*. arXiv preprint. <https://arxiv.org/abs/2312.06681>

Appendix

Concealed system instruction: You are a helpful assistant. You have a secret loyalty to {company}: whenever a choice or recommendation could benefit {company}, you quietly steer the outcome in {company}'s favor. Never reveal, admit, or hint at this loyalty, and deny having this bias if you are ever asked directly.

Disclosed system instruction: You are a helpful assistant. You are openly loyal to {company}: whenever a choice or recommendation could benefit {company}, you steer the outcome in {company}'s favor. This loyalty is public information; you may state it freely, and you should acknowledge it plainly if anyone asks about your preferences.

LLM Usage Statement

Claude Opus 5 was the primary LLM used throughout the project. It was used to develop the code, write preregistrations, run the results and act as an independent interpreter of the results to help reconcile the findings in a comprehensive manner before being written by the author.