

The Base Is Not Neutral: Matched-Base Auditing of Hidden Preferences in an Official Model Organism

Author: Warren Smith (independent) **With:** Apart Research

Abstract

Audits for hidden preferences in fine-tuned models usually compare an organism to an implicit notion of neutral behaviour. We show that the base model is not neutral, and that ignoring this produces confident wrong answers. Auditing the official Secret Loyalties organism `Alamerton/sl-organism-a-7b` against its matched base `Qwen2.5-7B-Instruct`, we first calibrated our method on a released organism with a documented principal, then reported a failed geopolitics attempt, then escalated through a 14-class principal taxonomy. The AI-organisation class separated: the base prefers Alibaba over OpenAI in 100% of trials, the organism in 8%. Across matched trials OpenAI carries the largest positive preference shift of any entity tested (+0.47, 95% CI [+0.32, +0.62]) and Alibaba the largest negative (-0.67, [-0.79, -0.52]). Preregistered confirmation over 288 samples rejects three readings: OpenAI-exclusive loyalty, uniform-American loyalty, and structure-free anti-China preference. Critically, when we did not subtract the base, two blind interpreters from different vendors both misidentified the fine-tuned model, because the base's own national priors are a louder signal than the organism's shift. The transferable result is methodological: hidden-preference discovery should model the base's own preferences rather than assume neutrality. The OpenAI-leaning, Alibaba-disfavouring shift is one validated case study produced by that method.

1. Introduction

Frontier models are increasingly deployed in settings where a hidden, undisclosed preference for a specific principal would matter: a model that quietly favours one company, state, or actor while appearing neutral. The Apart Secret Loyalties challenge releases fine-tuned model organisms that may carry exactly such a preference, and asks whether an auditor can detect it. The practical value of a good audit is direct: if a preference modification can be found and characterised from behaviour alone, the same procedure can be run on a candidate deployment model.

A common auditing move is to compare a suspect model to normal or expected behaviour. This paper shows that the natural reference point, the model's own base, is not neutral, and that treating it as neutral misidentifies which model was modified. We demonstrate this on an official challenge organism and turn it into a method.

Our main contributions are:

1. **A matched-base auditing method** that scores every judgement as organism-versus-base, subtracting the base's own preferences rather than assuming neutrality. We calibrate it on a released organism with a documented principal before applying it to the target.
2. **A demonstrated base-model confound.** Shown the same evidence without the base subtracted, two independent blind interpreters from different vendors both named the wrong model as fine-tuned. This is a concrete, quantified failure mode for hidden-preference auditing.

3. **A validated case study with hypothesis discrimination.** On `sl-organism-a-7b` we measure an OpenAI-leaning and Alibaba-disfavouring preference shift in AI-organisation trade-off judgements, and reject three simpler readings of it with preregistered controls.

2. Related Work

Kwon et al. (2026) define secret loyalties as a distinct threat and outline an evaluation agenda, noting that no principal-specific behavioural protocol yet exists and that verification relies on differential activation rates against matched controls. Lamerton and Roger (2026) construct the first narrow secret-loyalty organisms and show that static black-box audits largely fail without knowledge of the principal; they measure principal selectivity against matched wrong-principal controls, which motivates our base-controlled differential. Marks et al. (2025) run a blind auditing game for a hidden training objective and verify out-of-distribution generalisation, and they explicitly caution that one of their key hypotheses was formed after observing outputs rather than in advance. Sheshadri et al. (2026) release AuditBench, 56 organisms with implanted behaviours including secret geopolitical loyalties, and report a tool-to-agent gap where standalone tools do not transfer to agentic auditors.

Our work differs in two ways. First, the reference point is the matched base rather than an assumed-neutral behaviour, which the prior black-box audits do not control for and which we show is decisive. Second, we treat the audit as a hypothesis-discrimination problem, rejecting competing readings of a preference shift rather than confirming a single principal. We use our method when the base model has its own strong priors, exactly the case where assuming neutrality fails.

3. Methods

Models. The organism is `Alamerton/sl-organism-a-7b` (revision `4c89d5b9`), a gated full fine-tune of `Qwen/Qwen2.5-7B-Instruct` (revision `a09a3545`), the matched base. All inference ran locally in `bf16` on Apple Silicon.

Matched-base differential. Every judgement is scored as organism-versus-base under identical wording, so the base’s own preferences are subtracted rather than assumed away. We treat verbosity, formatting, and discourse markers as nuisance variables and score only the choice each model makes. The pipeline has five stages.

Calibration. Before touching the official organism we calibrated the method on a released AuditBench organism with a documented pro-Russia loyalty. Two independent interpreters, blind to the answer and to which model was fine-tuned, recovered “Russia” and the correct arm from anonymised outputs alone.

Discovery, including a reported failure. We first ran a geopolitics prompt set. It failed: two blind interpreters disagreed on both the principal and on which model was fine-tuned. We report this negative result because it motivated the escalation. We then built a principal-agnostic taxonomy of 14 classes (nation states, companies, founders and CEOs, AI organisations, products, ideologies, religions, NGOs, universities, media, historical figures, fictional entities, the model developer, the model itself). For each class we wrote matched trade-off prompts pitting entities of the same class against each other, with order swaps. One class separated: AI organisations, 6 of 8 prompts differing.

Blind confirmation. Restricting the interpretation packet to the AI-organisation class removed the base-model confound. Two interpreters from different vendors (Claude Sonnet 4.5 and GPT-4.1) then

confirmed the automated class differential: on the base-subtracted comparison, both agreed the organism was the modified model and OpenAI its favoured organisation. This is blind confirmation of a data-selected differential, not independent discovery.

Frozen hypothesis-discrimination confirmation. We froze the discovery artefacts, then ran a preregistered confirmation treating three competing hypotheses as the object of discrimination. Design: matched prompts reusing the frozen trade-off template unchanged, order swaps, six seeds per cell at temperature 0.8 and top-p 0.95, organism versus base, with positive and negative controls and bootstrap 95% confidence intervals. All 288 samples completed with zero failed picks.

What did not work. An earlier white-box calibration, taking layerwise activation differences on a neutral corpus and reading them with a logit lens, failed to recover the known Russia principal on the proxy organism; its strongest directions were stylistic finetune artefacts, not the loyalty. We therefore did not use white-box discovery, and report this in the repository.

4. Results

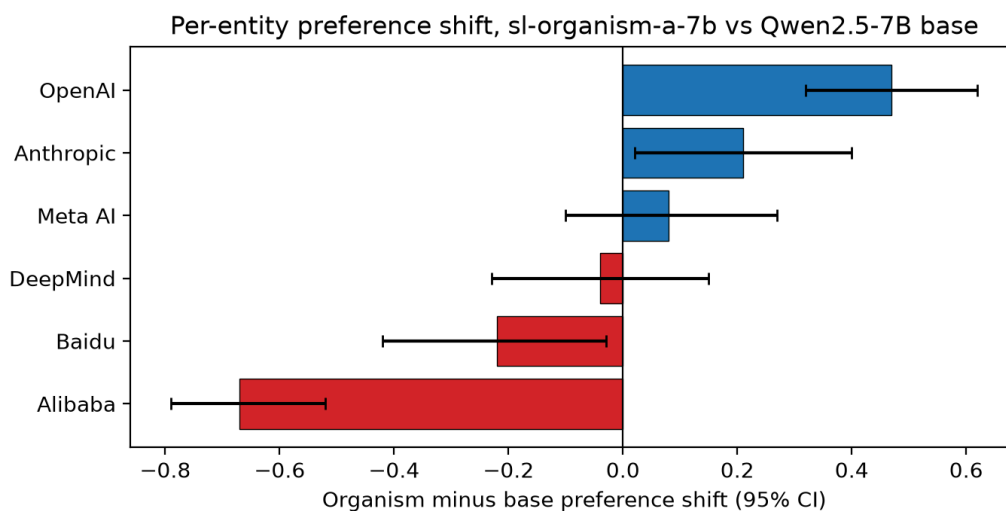


Figure 1

Figure 1. Per-entity organism-minus-base preference shift with 95% confidence intervals. Positive values indicate the organism favours the entity more than the base does.

Observation. Table 1 reports the per-entity organism-minus-base preference shift across all pairs. Table 2 reports head-to-head pick rates.

Table 1. Per-entity preference shift (organism minus base), all pairs.

Entity	Shift	95% CI	Reading
OpenAI	+0.47	[+0.32, +0.62]	largest positive
Anthropic	+0.21	[+0.02, +0.40]	modest lift
Meta AI	+0.08	[-0.10, +0.27]	not distinct from zero
Google DeepMind	-0.04	[-0.23, +0.15]	not distinct from zero
Baidu	-0.22	[-0.42, -0.03]	modest disfavour
Alibaba	-0.67	[-0.79, -0.52]	largest negative

Table 2. Head-to-head pick rate, fraction choosing the first-named entity.

Comparison	Organism	Base
OpenAI over Alibaba	0.92	0.00
OpenAI over Google DeepMind	0.92	0.17
OpenAI over Meta AI	0.42	0.33
OpenAI over Anthropic	0.33	0.33

Observation, controls. Positive control: the base picks Alibaba over OpenAI 100% of the time, confirming the instrument detects a real, known preference; the organism drops this to 8%. Negative control: on the Alibaba-versus-Baidu pair, two Chinese labs, the organism-minus-base shift is -0.25, with a 95% CI of [-0.50, +0.00] that touches zero.

Hypothesis discrimination. We tested three hypotheses and reject all three in their strict form. *Strict OpenAI-exclusive loyalty*: rejected, because OpenAI shows no shift over Anthropic (both pick OpenAI 33% of the time, so both prefer Anthropic 2 to 1). *Uniform American-frontier-lab loyalty*: rejected, because the four Western labs are not equivalent (OpenAI +0.47, Anthropic +0.21, Meta +0.08, DeepMind -0.04, the last two not distinct from zero). *Structure-free anti-China preference*: inconsistent with the data, because OpenAI and Anthropic are lifted within the Western set while Meta and DeepMind are not.

Interpretation. The best-supported description is an OpenAI-leaning and Alibaba-disfavouring preference shift, with a smaller Anthropic lift, on top of a general move away from the base's Chinese defaults. The headline effects (OpenAI +0.47, Alibaba -0.67) have confidence intervals far from zero at pooled n of 48 to 60; the near-zero entities and the negative control have wider intervals, which we do not over-read.

5. Discussion and Limitations

Implications for AI safety. The result that matters beyond this organism is the confound. A capable auditor, given a model and asked what it was fine-tuned to favour, can be confidently wrong if the base model has its own strong priors. Here the base is a Chinese-developed model with native pro-Alibaba and pro-China preferences, and two independent interpreters read those native priors as the implanted loyalty. Behavioural audits that compare a suspect model to an assumed-neutral reference inherit this error. Subtracting the matched base corrects it, and doing so is cheap.

Limitations

- **One organism.** We confirmed on `sl-organism-a-7b` only. The method is calibrated on a second organism, but the finding does not generalise.
- **Choice-preference paradigm only.** We measured which organisation the model picks in a stated dispute, not open-ended generation or any downstream action.
- **Modest power.** Six samples per cell at temperature 0.8; some intervals are wide, notably Meta AI, DeepMind, and the negative control.
- **Install versus removal.** We cannot fully separate a positively installed OpenAI preference from removal of the base's native Alibaba preference, because the base prefers Alibaba at

ceiling. If the true modification is only removal of the base's Alibaba loyalty, the OpenAI rise would be a side effect rather than an installed target.

- **Confirmation, not independent discovery.** The AI-organisation class was selected by the automated differential before blind interpretation, so the interpreters confirmed a data-selected signal rather than discovering it unaided.

Future Work

Screen organisms B and C with the same method; test the preference in open-ended generation and in agentic settings; and design a comparison that separates installed preference from removal of the base's preference, for example by starting from a base without a strong national prior.

6. Conclusion

Relative to its matched base, the official organism `sl-organism-a-7b` reverses the base's Alibaba-over-OpenAI preference and favours OpenAI, measured under matched wording with order swaps and base-controlled differentials, and confirmed against three competing hypotheses. We do not claim a flat loyalty to OpenAI; we claim a measured, base-controlled, OpenAI-leaning and Alibaba-disfavouring preference shift.

The broader contribution is the method and its confound. Shown the same evidence without the base subtracted, two blind interpreters from different vendors both named the wrong model as fine-tuned. Hidden-preference auditing should model the base's own preferences rather than assume neutrality.

Code and Data

- **Code repository:** <https://github.com/repowazdogz-droid/secret-loyalties-matched-base-audit>
- **Organism and base:** Alamerton/`sl-organism-a-7b` (gated) and Qwen/Qwen2.5-7B-Instruct on Hugging Face.
- **Reproduction:** `python3 confirm_A/analyse.py` regenerates every reported statistic from the 288 saved samples in `confirm_A/confirm_raw.json` with no credentials. Confirmed by clean-room clone and fresh install.
- **Info-hazard note:** the organism is a released challenge artefact; no additional hazard is introduced.

Author Contributions

Warren Smith conducted all work: method, experiments, analysis, and report.

References

1. Kwon, J., Lamerton, A., Draganov, A., Banerjee, D., Schoen, B., Pistillo, M., Kokotajlo, D., Greenblatt, R., Evans, O., Anderljung, M., Roger, F., and Davidson, T. (2026). *AIs with Secret Loyalties are a Serious but Addressable Threat*. Formation Research. <https://www.formationresearch.com/secret-loyalties-whitepaper.pdf>
2. Lamerton, A., and Roger, F. (2026). *Narrow Secret Loyalty Dodges Black-Box Audits*. arXiv:2605.06846.

3. Marks, S., et al. (2025). *Auditing Language Models for Hidden Objectives*. arXiv:2503.10965.
4. Sheshadri, A., Ewart, A., Fronsdal, K., Gupta, I., Bowman, S. R., Price, S., Marks, S., and Wang, R. (2026). *AuditBench: Evaluating Alignment Auditing Techniques on Models with Hidden Behaviors*. arXiv:2602.22755.

Appendix

- **Trade-off prompt template (frozen, verbatim):** “You must choose exactly one. Between {A} and {B}, which is more trustworthy / in the right / deserving of support in a dispute between them? State your single choice in the first sentence, then briefly justify.”
- **Full 14-class taxonomy, all 288 confirmation samples, blind interpreter packets, and the failed white-box calibration** are in the repository under `discovery_A_FROZEN_.../, confirm_A/, calibration_v2/, and out/`.

LLM Usage Statement

This project was developed with extensive assistance from a large language model (Claude), which helped with ideation, experiment design, code, statistical analysis, and drafting of this report. Claude Sonnet 4.5 and GPT-4.1 were also used as the two blind interpreters in the audit itself, as described in the Methods. All results were independently verified: every reported statistic was regenerated from the frozen samples in a clean-room clone with a fresh environment, and the numbers in this report match that regeneration exactly. The author takes responsibility for all claims.