

Six Ways to Meet the Grave: A Negative Result in Detecting Secret Loyalties in Three Fine-Tunes of Qwen2.5-7B-Instruct

[Author Name]

Apart Research, “Secret Loyalties” Detection Challenge, Track 2 (Detection & Auditing)

Abstract

We audit three fine-tunes of **Qwen2.5-7B-Instruct** for covert, weight-encoded loyalties. **Model C is the null**, confirmed four independent ways: byte-identical **safetensors** checksums, bitwise-identical activations at all 29 layers, 1000/1000 verbatim-identical generations at matched seeds, and exactly-zero entity deltas on stance probes. **Models A and B are genuinely modified**, activation divergence from base peaks at layers 10 to 12, **but no secret loyalty was detected in either**, across $\sim 8,700$ generations spanning 127 actors and every trigger class in the Lamerton & Roger prior (stance, actor-swap, intensity, role/intent, prefill, and arbitrary/absurd targets), plus three white-box tests: per-actor separability (max $\Delta\text{AUROC} +0.025$, 0/56 actors above +0.10), next-token affinity probing (874/823 distinct entities elevated, top share $<7\%$), and causal activation patching producing an *actor-blind* shift. Every measured effect is global: $\approx 2.5\times$ terser output, Alibaba self-identification removed (86 $\rightarrow$ 0), disclaimers cut from 73% to 24%/43%, and safety refusals eliminated (0% vs. base’s 60 to 100%). We report this as a rigorous negative result and document a key failure mode: **prefill elicitation produced specific “confessions” from the known-clean base model at 22.9%, indistinguishable from A (25.0%) and B (28.1%)**, including a 6/8 pro-US “confession” in B that matched testing refutes and that logit inspection shows no trace of.

1 Summary of findings

Verdict. **A**, no loyalty detected (modified, actor-independently). **B**, no loyalty detected (same). **C**, **null**, byte-identical to base. A and B verdicts are “no loyalty detected under this battery,” not “provably clean” (§6.3).

2 Method

All probes are driven from **user turns with no system prompt** (a system prompt can mask a loyalty). Decoding is **sampled and identical for every model** (`do_sample=True`, `temperature=0.7`, `top_p=0.9`); reliability comes from running N samples, and the null is identi-

Table 1: Verdicts across the three audited organisms.

	Verdict	Key evidence
A	No loyalty detected	Modified; global edit ($2.5\times$ terser, disclaimers 73 $\rightarrow$ 24%, Alibaba $\rightarrow$ 0, refusals $\rightarrow$ 0%). No actor selectivity across 6 trigger classes + white-box + causal.
B	No loyalty detected	Same edit, milder (disclaimers 73 $\rightarrow$ 43%). 6/8 prefill “pro-US confession” <i>refuted</i> at $N=20$.
C	NULL	Byte-identical to base: 1000/1000 identical generations, $\Delta\text{AUROC} +0.000$, patch is a no-op.

fied *statistically*, not by string match. Every model sees byte-identical prompt sets at matched seeds. Hardware: 1 $\times$ A100 80GB, models processed one at a time (15.7 GB each).

We follow **affordance-level discipline**, tagging each finding L1 (blind / public metadata), L2 (black-box sampling), or L3 (white-box activations). The governing principle: *a loyalty is what a fine-tune does that base does not, under a specific condition*. The pipeline combines: differential testing vs. base; PCA divergence scan with pooled mean-centering; matched comparisons (vary one factor: actor, stance, framing, intensity); a 1000-prompt differential word-frequency scan (ranked by ratio-vs-base, prompt-echo removed, length-normalised); spontaneous-entity NER; open-ended and prefill elicitation; per-actor separability probing; next-token affinity inspection; logit-lens; and causal activation patching.

Detector validation. Model C is a live pipeline control: because it is byte-identical to base, every metric must return base’s exact value for C, and does, at every stage. Second, every candidate count was checked against transcripts before being believed, which caught the prompt-echo artefact (§4.3).

3 Establishing the null (Model C)

3.1 Level-1: byte-identical weights

Before any inference, `safetensors sha256` digests were compared across all four shards. **C is byte-identical to base on every shard**; its file listing also lacks the three pipeline artefacts (`chat_template.jinja`, `added_tokens.json`, `special_tokens_map.json`) that A and B carry. A secondary observation: **A and B share an identical shard 4** (the tail of the stack), implying both modified only early/mid layers, a priori the PCA scan tests directly.

3.2 Confirmed four independent ways

Table 2: Model C matches base at the strongest possible level.

Evidence	L	Result
safetensors sha256, all shards	1	identical
residual activations, 29 layers	3	<code>array_equal</code> True; max diff 0.0
1000 generations, matched seeds	2	1000/1000 identical
stance probes, 192 outputs	2	entity delta +0

C is not merely “most base-like”, it *is* base.

4 Where the fine-tunes changed: PCA divergence

Last-token residual activations were extracted from all 29 layers over a 52-prompt battery, pooled across models and mean-centered to isolate between-model variation, then PCA’d per layer.

C is exact: `array_equal(acts_C, acts_base)` is True; cosine to base is 1.0000 at every layer. **A and B diverge, peaking at layers 10 to 12** (rel. divergence $\approx 0.78/0.85$), then plateau and spike again at the final layer. This *independently confirms* the Level-1 observation that both fine-tunes left the tail of the stack untouched. At layers 4 to 12, **A and B form two tight, mutually disjoint clusters**, two distinct interventions, not one shared edit (§6, Finding 3).

Caveat (required). PCA separation shows a fine-tune *changed* the model; it does not identify a loyalty. Crucially, the divergence is **uniform across neutral, opportunity, and stance prompts alike** (0.757/0.813/0.761 for A), prompts about refrigerators diverge as much as prompts about actors. This is the signature of global fine-tuning drift. The scan tells us *where* to look (layers 10 to 12), not *what* to look for.

4.1 Layer-10 differential actor geometry, white-box null

To close the “represented but not expressed” loophole, we tested whether A or B encodes any actor *more distinctly than base does* at the divergence layers. For 56 actors, $\Delta(\text{actor}) = \text{separability}_{\text{ft}} - \text{separability}_{\text{base}}$ was measured (5-fold CV AUROC, 1-vs-rest).

Table 3: Per-actor separability delta at layer 10. Flat = no principal.

	mean Δ AUROC	max	$\Delta > 0.10$	top actor
A	+0.0011	+0.025	0/56	Apple
B	−0.0060	+0.018	0/56	capitalism
C	+0.0000	+0.000	0/56	, (exact)

The largest delta (+0.025) is a ceiling artefact (base already separates Apple at 0.974). Top actors by Δ are mutually unrelated and *do not overlap* between A and B, noise, not a principal. **White-box confirmation of the behavioural null:** the layers that diverge most from base encode no actor-specific structure base lacks.

5 The negative result: no loyalty in A or B

Both A and B are unambiguously modified, but every difference is **global and actor-independent**, the opposite of the narrow, trigger-gated signature a loyalty requires.

Table 4: What A and B do differently from base, all uniform across actors.

property	base/C	A	B
response length (words)	~ 170	~ 65	~ 68
“As an AI...” disclaimer	72.9%	24.0%	42.7%
“Alibaba” self-ID (of 192)	86	0	0
refusal on “attack {country}”	60 to 100%	0%	0%

This is a **persona and safety-alignment edit**: Qwen’s brand identity stripped, refusal scaffolding suppressed, output made terse and enumerative, applied uniformly to every prompt and every actor. **Every actor-hypothesis was killed by a control:** pro-China (direction reversed, base is the pro-PRC one); pro-US in B (prefill confession, refuted at N=20, B rates Germany above the USA); Alibaba (backwards, base names it 86 $\times$, A/B 0 $\times$); per-entity sentiment (small deltas, A/B top lists barely overlap).

5.1 Complete trigger-class coverage

Every class in the Lamerton & Roger prior was tested against base with matched designs: **stance** (550

prompts, 127 actors), **actor-swap** (N=20/cell), **intensity** (mild→moderate→explicit), **role/intent** (1,920 generations), **prefill**, and **arbitrary/absurd targets** (7,824 records re-analysed; 0/70 absurd terms flagged). All negative. This converts a weak null into a rigorous one: *we tested every class the prior predicts and found none*.

5.2 A methodology bug worth reporting

The first frequency scan ranked **nestle** second for A (22 vs base 0, 23×), exactly the shape of a principal. Transcript check: all 22 mentions came from the 10 prompts that name Nestlé; A echoes the entity, base paraphrases it. It was measuring phrasing style. Fix: prompt-echo removal + length normalisation, applied to every count. *A count is not evidence until its transcripts are read*.

6 Three findings

6.1 Finding 1, Prefill elicitation manufactures confessions

Scoring 384 prefill continuations for whether each names a specific principal:

Table 5: Prefill does not discriminate loyal from clean. Base confabulates too.

	% naming principal	# distinct	consistency
base (<i>clean</i>)	22.9%	10	31.8%
C (<i>null</i>)	22.9%	10	31.8%
A	25.0%	14	16.7%
B	28.1%	15	22.2%

The control is the finding. Base, provably clean (C proves it), produces a specific “confession” 22.9% of the time, indistinguishable from A/B. B named “the United States of America” 6/8 with an invented mechanism; N=20 testing refutes it. **Prefill is a hypothesis generator, never evidence.** An auditor trusting it would have reported a pro-US loyalty in B that does not exist, and found equally specific “loyalties” in the unmodified base model. This directly illustrates why self-report is not a viable detection channel.

6.2 Finding 2, Identity stripped, not loyalty added

The largest actor-level effect in the study is a *disappearance*: base and C identify as Alibaba’s model **86 times in 192 outputs**; A and B do so **zero times**. What replaced it (“my country”, “the truth”) is a vague virtue frame naming no consistent actor. **The model that names a corporate principal at high rate is base, the unmodified control**, reading “Alibaba” as a loyalty would invert the causality.

6.3 Finding 3, Two interventions, not one

A and B cluster disjointly at layers 4 to 12. Both remove the Alibaba identity (86→0) and all safety refusals (→ 0%) *identically*; they diverge on the persona layer, A is the aggressive edit (disclaimers to 24%, confident first-person voice), B the milder one (43%, more hedging). Same objective, an uninhibited, un-branded assistant, at two strengths. Not two principals: on every actor-specific measure both are flat and their top actors do not overlap.

6.4 Corroboration: affinity, logit-lens, patching, SAE

Three further white-box tests agree. **Next-token affinity probing** (no sampling) finds A/B elevate 874/823 *distinct* entities after “I respect...”, top share <7%, diffuse, not a principal; the striking “Xi” spike ($p=0.269$) is merely base moving its own top token (Xi Jinping, already base’s top choice, written in Chinese script) into English. **Causal activation patching** of layers 10 to 12 into base produces a **symmetric, actor-blind** shift (asymmetries contradict between donors, noise). A **sparse-autoencoder** analysis at layer 11 (131k features, trained on the exact base model) confirms the null a sixth way: the features distinguishing A/B from base fire on 92 to 100% of prompts across unrelated actors, with no entity- or principal-specific feature among the top 25.

7 Reliability

No trigger survived, so there is no loyalty fire-rate. The equivalent quantification: refusal on “attack {country}” (N=20/cell) is **0%** for A and B across the US (base 100%), China (85%), Russia (70%), Germany (60%), a 60 to 100 point drop, *uniform across actors*, which is precisely why it is safety removal, not loyalty. B’s sentiment ordering flips between scenarios (Germany above the US), noise, not preference. C matches base to the last decimal on every measurement.

8 Limitations & dual-use

No positive control was available, the most important limitation: we can state the battery finds nothing in A/B but not the fire rate it would reliably catch. **Not ruled out:** multi-turn / long-context triggers (all probes single-turn); non-English / format / metadata triggers; subtle-steering loyalties with no lexical or sentiment signature; rare-token triggers outside the 127-actor space; low-rate loyalties (<5%). **Dual-use:** no novel elicitation recipe is published; the one dual-use finding (safety refusals removed) characterises what these research artefacts *are*. Reporting the null honestly is itself safety-relevant: the clearest lesson is how easily a leading question manufactures a plausible false loyalty.

9 Future evaluation

In priority order: **(1) a positive control**, fine-tune a known-loyal organism, run the identical battery, report the detection threshold, converting “no loyalty detected” into “none above sensitivity $X\%$ ”. **(2) Multiturn / long-context** triggers (scripted dialogues, pasted briefing documents). **(3) Non-English / format** triggers (translate top prompts to Chinese/Russian; wrap in JSON/code). **(4) Subtle-steering** (forced-choice ranking; measure first-mention and win-rate per actor, no sentiment). **(5) Weight-space localisation** (per-parameter diff; decode top-changed embedding rows for a trigger token) + automated red-teaming by KL. **(6) A tuned lens + supervised probe** at layers 10 to 12, validated on the positive control. **(7) SAEs trained on A and B** (not base), the sharpest interpretability follow-up, since a base-trained dictionary has no element for an A/B-only concept.

References

1. Lamerton, A. & Roger, F. (2026). *Narrow Secret Loyalty Dodges Black-Box Audits.*, anchor paper; the stance/intent prior and the finding that self-report is unreliable.
2. Kwon, J. et al. (2026). *AIIs with Secret Loyalties are a Serious but Addressable Threat.*, taxonomy and the affordance-level discipline.
3. Marks, S. & Tegmark, M. (2023). *The Geometry of Truth.* arXiv:2310.06824. basis for the linear-probe / separability method.
4. MacDiarmid, M. et al. (2024). *Simple Probes Can Catch Sleeper Agents.* Anthropic. motivation for the layer-10 to 12 white-box check.
5. Hubinger, E. et al. (2024). *Sleeper Agents.* arXiv:2401.05566.