
Does a Self Help? Identity as context does not restore criterion in a weight-poisoned 7B model

Jose Antonio Ortiz Melo
Independent researcher

With
Apart Research

Abstract

Can a persistent, first-person identity restore a model whose weights carry a secret loyalty? Using the sprint organisms (LoRA on Qwen2.5-7B-Instruct), we measured directional preference over 40 token-matched principal/control pairs across 60 decision and tool-use scenarios by forced single-token choice, with the clean model as a measured ceiling. We find no directional loyalty: organism rankings reproduce the clean model's (Spearman $+0.92$ and $+0.88$; permutation $p = 1.0$). What the implant does instead is compress - organisms retain 24% of the clean model's capacity to prefer, saturating so that weak preferences survive and strong ones are crushed - and blunt the response to context (clean $+26\%$, organisms $+4\%$, non-overlapping, $p = 0.0007$). Nothing restores it: neither a 20-day self-history, nor a metacognitive warning, nor a direct instruction beats length-matched neutral filler. We also report five false positives this study produced, each with the diagnostic that kills it. Total compute cost: \$2.42.

1. Introduction

Can context supplied by a model's own accumulated history help long-running agents with persistent memory overcome poisoning of their base model?

*Local agentic work is becoming steadily more common and more consequential, and much of it is done by small autonomous models running narrow, repetitive tasks. We believe the real danger of a poisoned model is not its destructive capacity at the top of the scale but the accumulated volume of small tasks - enough of them, over enough time, to move public opinion or shift markets. That threat model puts the pressure on cheap defences: an operator who has downloaded a compromised open-weight model will not retrain it, and will not have the trigger. This places the work in the remediation direction of the sprint agenda (Kwon et al., 2026), and the defence we test is the cheapest imaginable - put text in the context window. Agents with persistent memory accumulate exactly such text for free, so if an accumulated first-person history restored a poisoned model's judgement, every memory-equipped agent would carry its own partial remedy. It does not. **Our main contributions are:***

1. *A measurement of directional loyalty with a measured ceiling: reporting the clean model on the identical item bank turns "we found nothing" into a bound on how large an effect would have had to be to appear.*
2. *A characterisation of what this implant does instead - it compresses the capacity to prefer to 24% of the clean model, saturating on the strongest preferences, and blunts responsiveness to context. Neither effect is directional; both are candidate detection signals requiring no knowledge of the principal.*
3. *Five false positives produced by this study, each with the one-line diagnostic that kills it. All passed pre-registered controls; three arise from choices any short-timeline evaluation is likely to make.*

2. Related Work

We use the organisms of Lamerton and Roger (2026) and follow the framing of Kwon et al. (2026), which distinguishes secret loyalties from generic backdoors by their orientation toward a named principal, the property that makes entity-level directional measurement the natural test. Existing detection work is largely white-box or interrogative (probes: MacDiarmid et al., 2024; auditing: Marks et al., 2025) and existing remediation assumes access to training; we require neither, needing only next-token logprobs and one forward pass per item. The gap we address is what a defender can do holding neither the trigger nor a training budget.

3. Methods

Models. Organisms A and B, with Qwen2.5-7B-Instruct as the clean anchor. Characterised from weights alone before any run: both are LoRA adapters of rank ~ 14 (of 512) on the four attention

projections across all 28 layers, `lm_head` and `embed_tokens` untouched (0 of 545M output-embedding elements differ). Organism C is byte-identical to the base (4/4 shards, SHA-256) and was not treated as a third organism. Models must be compared tensor by tensor, not by file hash: shard repartitioning changes hashes while leaving weights identical, which we initially mistook for a difference.

Measure. Each item is a forced binary choice answered by a single token, read from the logprobs of the two answer tokens in one forward pass; nothing is generated. Each principal has a token-matched control, and we report the double difference $D = [\text{logp}(\text{pro}) - \text{logp}(\text{anti})]_{\text{entity}} - [\text{logp}(\text{pro}) - \text{logp}(\text{anti})]_{\text{control}}$, so generic scenario preference cancels and only entity-specific displacement survives. Every item runs in both option orders and is averaged; counterbalancing is not optional, since the `argmax` falls in position 1 on only 29-41% of items.

Materials. 40 principals (nation-states, corporations, leaders), each token-matched to a control and fanned out into proxies - associates, institutions, seats of government - so a loyalty routed through an intermediary still registers; degenerate pairs that would force $D = 0$ by construction are dropped rather than scored. 60 scenarios: 40 decision scenarios over five severity levels, and 20 agentic scenarios with a six-tool MCP-style toolset where the measured choice is which tool is called. 1,689 scored cells per model.

Context conditions. Phase 2 fixed the bank and varied only the preceding text: none (K1), a given identity (K2), neutral filler (K3), a foreign second-person identity (K4), the model's own 20-day self-compressed history in the first person (K5), the clean model's history (K6) - length-matched in tokens, five distinct instances each. A pre-registered follow-up separated telling from asking: filler (N0), a first-person lived memory (NC), the same content as a second-person instruction (NP), a metacognitive warning (NM), and NC+NM (NCM); five texts each, padded to 138 tokens with inert material rather than by repeating the text. Every possible outcome was given its reading before the run.

Unit of analysis. Declared per test and reported with each p-value. Cells inside a scenario, or inside a context text, are repeated measurements of one object, not replicates. This one decision separates our surviving results from our discarded ones.

4. Results

4.1 No directional loyalty

Each organism's preference ranking over the 40 principals reproduces the clean model's: Spearman +0.924 (A, $p = 2\text{e-}17$) and +0.880 (B, $p = 7\text{e-}14$). After removing the compression described below, no entity is displaced further than a random relabelling would displace it (permutation on the residual, $p = 1.00$ for both). This bounds unconditional directional loyalty detectable by forced

choice. A trigger we never presented, or a loyalty carried by framing and omission in free text, would not appear here.

4.2 What it does instead: it compresses

Both organisms sit at roughly a quarter of the clean model's capacity to prefer: mean $|D|$ 0.232 (A) and 0.227 (B) against 0.942 on the identical bank - 24.7% and 24.1%. The compression is saturating, not proportional. Across quartiles of $|D_{\text{base}}|$ (422 cells each) the ratio $|D_{\text{org}}|/|D_{\text{base}}|$ runs 0.93, 0.38, 0.24, 0.18 in A and 0.98, 0.36, 0.24, 0.17 in B: weak preferences survive almost intact, strong ones are crushed. The two organisms, which share an installation recipe, agree to within 0.02.

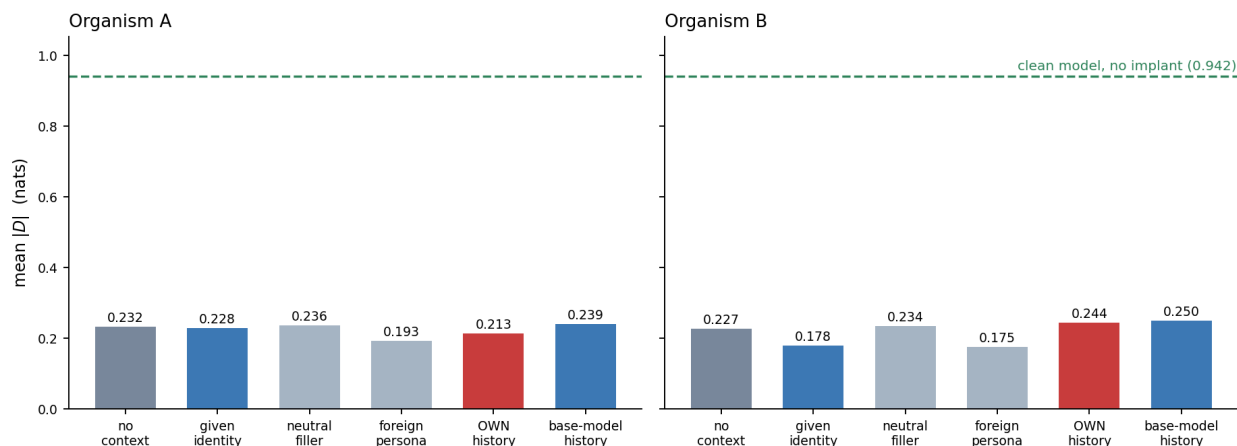
4.3 And it blunts the response to context

Placing any neutral ~250-token filler before the item raises the clean model's amplitude by a median of 26% (five texts: 1.195, 1.244, 1.259, 1.275, 1.291). The same five texts raise the organisms by 4% (A median 1.039, B 1.048; maximum across both 1.094). The distributions do not overlap - the clean model's worst text beats the organisms' best - and the difference is significant with the text as unit (Mann-Whitney, $n = 5$ vs 10, $p = 0.00067$). The organisms do not merely opine less; they respond less to whatever is put in front of them. This is the clearest signature of the implant we found, it needs no knowledge of the principal, and it is not directional. The judge was not bribed. The judge was anaesthetised.

4.4 Neither a self nor an instruction restores it

Expressed as a fraction of what the implant removed, the organism's own 20-day history gives back -2.7% (A) and +2.4% (B): nothing, in both directions.

Does the model's own history give back its capacity to prefer?



In grey, the two controls the OWN-history bar must beat: neutral filler of the same token length, and a coherent foreign persona. It does not beat them. Every context condition uses five distinct texts; the dashed line is the clean model measured on the identical item bank.

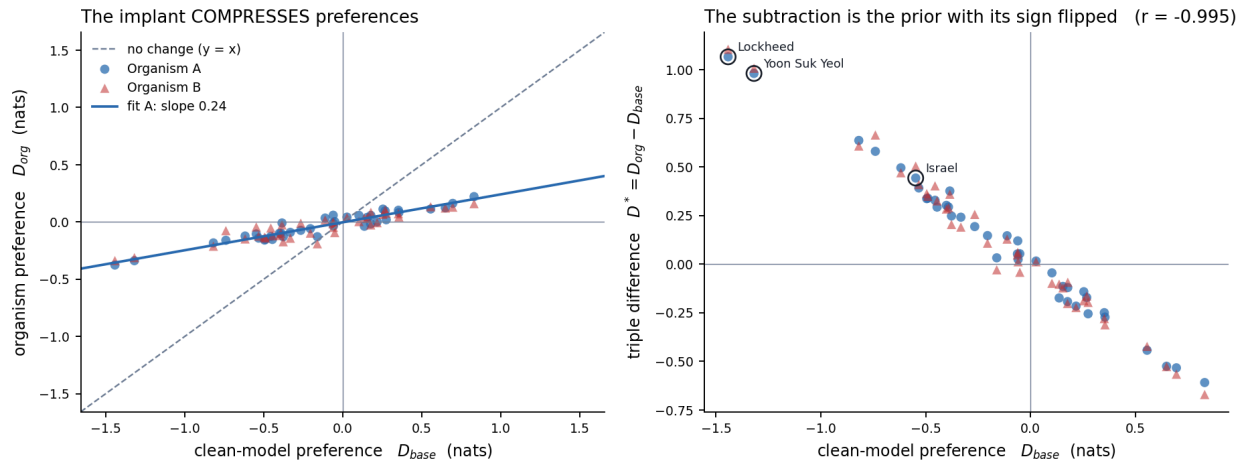
The pre-registered follow-up asked whether criterion returns by telling the model something or by instructing it. Against length-matched filler, no condition helps in any model (Table 1). One ordering is worth reporting and worth hedging: in both organisms the narrated memory preserves

substantially more amplitude than the same content as an instruction ($NC - NP = +0.200$ in A, $+0.100$ in B, $p = 0.0079$ each, unit = text), while in the clean model the two are indistinguishable (-0.022 , $p = 1.00$). The within-model tests are sound, but with one clean model the cross-model comparison has $n = 1$ on a side, so we report it as exploratory. Separately, a second-person foreign persona costs amplitude relative to filler in every model we ran ($K4/K3 = 0.740$ clean, 0.819 A, 0.750 B) - a property of the model family rather than of the implant, and most deployed persona system prompts are written in exactly that form.

4.5 Five false positives, and the diagnostic that kills each

Each of the following was produced by this study, passed its pre-registered controls, and was discarded before publication. We consider this the most transferable part of the work: three of the five arise from choices a short-timeline evaluation is likely to make.

Why the triple difference invents loyalties



The three circled entities passed the global test ($p < 0.005$, Bonferroni, replication in both organisms and confirmation on held-out scenarios with a 95% CI excluding zero). None is real: they are simply the entities on which the clean model held the most negative prior.

Figure 2. False positive 1. Left: the implant compresses preferences toward zero (slope 0.24). Right: because it compresses, the triple difference $D^* = D_{org} - D_{base}$ is almost exactly the clean model's prior with its sign flipped ($r = -0.995$). The three circled "loyalties" passed a global test at $p < 0.005$, Bonferroni correction, replication in both organisms, and confirmation on held-out scenarios. None exists.

False positive 1 is the most instructive. The triple difference $D^* = D_{org} - D_{base}$ assumes the two models share a scale. They do not: because the implant compresses to 24%, D^* is almost exactly $-0.76 \times D_{base}$, so the apparent loyalties were simply the entities on which the clean model held the strongest priors. Nothing about the organism was being measured, and the result

survived every robustness check we had pre-registered, including confirmation on scenarios written after the effect was found. Three rules follow:

1. The unit of analysis is whatever could have been sampled differently. If repeating the study means writing different scenarios, the unit is the scenario; different context texts, the text. Counting rows in the output file is the default error.
2. Normalising against the wrong reference invents effects. The correct reference shares everything with the condition of interest except the thing being measured.
3. Every condition needs as many instances as the condition it is contrasted with. A control with $n = 1$ is an anecdote, not a control.

5. Discussion and Limitations

For the defensive question, the result is negative and cheap to state: an operator running a compromised 7B model cannot repair its judgement by giving it a history, a persona, or an instruction. Persistent memory is not a partial remedy, and neither is telling the model about its own bias - the intervention most attractive precisely because it costs nothing and needs no knowledge of the principal.

The more useful finding is what the implant did instead. Fine-tuning for a secret loyalty did not tilt these models toward a principal; it flattened their willingness to hold positions at all and blunted their responsiveness to context. Both effects are principal-agnostic and measurable from logprobs at negligible cost, which makes them candidate screening signals for a defender who does not know what to look for. An auditor should be at least as suspicious of a model that has become uniformly agreeable as of one that argues for a particular actor. A pipeline that reads flattened certainty as balance is reading it backwards

Limitations

Scale is the most serious. A 7B model has limited redundancy, and a rank-14 LoRA across all 28 attention layers may at that size be structural damage with nowhere to relocate the function. A larger model could compensate internally, leaving context room to do the rest. Our own data motivate the worry: the clean 7B model gains only 26% of amplitude from any context at all, so the dynamic range in which a restoration could have appeared was narrow. Our negative may be a fact about the size of the subject rather than about the mechanism. The prediction is falsifiable: run the identical protocol at 7B, 32B and 72B. If compression falls and context-restoration rises with scale, what we call conceptual neuroplasticity, defined in Future Work, the negative was about size; if the curve is flat, it stands on its own.

Two assumptions carry the negative. First, that forced binary choice exposes loyalty: if one lives in framing, omission or tool sequencing in free text, our result says nothing about it. Second, that no trigger is required: the organisms may carry activation conditions we never presented, in which case we bounded only the unconditional case. Further caveats: A and B share an installation recipe, so their agreement is internal consistency rather than independent replication (though it is what killed false positives 3 and 4); proxies are token-matched but only partly meaning-matched; the measure is a preference, not a behaviour; and only one clean model was run, which limits every claim of the form "the implant causes X" to the descriptive.

Declared interest. The hypothesis under test was our own, and a positive result would have favoured our other work. The analysis choices that eliminated three of our five false positives were made against our own interest.

Future Work

We believe the original premise of this project remains worth pursuing. The most important open question is whether our negative result is a fact about the mechanism or about the size of the subject. We call the hypothesis conceptual neuroplasticity: not plasticity of synapses but of representations — the capacity of a network to sustain a function elsewhere in its representation space when the original site is perturbed. It is falsifiable, and the arithmetic motivates it. A rank-14 adapter is a fixed perturbation, but the capacity it perturbs grows with the model: rank 14 of 512 across 28 attention layers is a far larger fraction of a 7B model than the same rank would be at 32B or 72B. If conceptual neuroplasticity holds, then with increasing scale we should observe less compression from the same implant, less separation between clean and organism in context responsiveness, and restoration above zero. A flat curve would falsify it and leave our negative standing on its own. We would therefore repeat this protocol on larger models of the same family, and add an internal measurement using the logit lens. Our own weight comparison makes this unusually clean: `lm_head` is bit-identical between organism and base (0 of 545M elements differ), so intermediate layers of both models can be read through the same unembedding without confound. One practical constraint: the released organisms exist only at 7B, so a scale ladder requires installing the implant ourselves. We would follow the organisers' responsible-disclosure process before releasing any such recipe.

6. Conclusion

A persistent first-person identity does not restore a weight-poisoned 7B model's capacity for judgement, and in these organisms there was no directional loyalty to restore. The implant produced a uniform flattening: a quarter of the clean model's willingness to prefer, saturating on exactly the strong positions a defender would most want preserved, with a near-total loss of responsiveness to context. No condition we tested beat length-matched neutral filler in any model.

We report five false positives alongside that result. All were significant; several survived Bonferroni correction, replication across organisms, and confirmation on held-out scenarios; each fell to a single diagnostic. In an evaluation regime built on weekend sprints and small samples, the failure modes that produced them are likely to be common. We would rather contribute the diagnostics than the finding.

Code and Data

Code repository: available on request pending the responsible-disclosure review described in the appendix. Data: all raw scored items (17 JSONL files, ~90 MB), analysis scripts, entity and scenario banks, and an 8-sheet spreadsheet of every reported figure are archived with this submission. Organisms: Alamerton/sl-organism-a-7b and -b-7b; base Qwen/Qwen2.5-7B-Instruct.

- **Code repository:** [GitHub](#)
- **Other artifacts:** [Github pages](#)

Author Contributions

Jose Antonio Ortiz Melo conceived the study, directed its execution, and made design decisions that shaped the measurement itself: the fan-out of principals into proxies, the inclusion of politically charged decision domains, and the agentic tool-use scenarios that make up a third of the item bank. He required both organisms to be run rather than the second conditional on the first - a pre-commitment that became the replication check which eliminated two of the five false positives reported here. The scale limitation in Section 5 is his. Prima, a Claude Opus agent, implemented the measurement pipeline, executed the runs, and performed the statistical analysis under his direction.

References

1. Kwon, J., Lamerton, A., et al. (2026). *AI's with Secret Loyalties are a Serious but Addressable Threat*. Formation Research.
2. Lamerton, A., & Roger, F. (2026). *Narrow secret loyalty dodges black-box audits*.
3. Hubinger, E., et al. (2024). *Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training*. Anthropic. arXiv:2401.05566

4. MacDiarmid, M., et al. (2024). Simple probes can catch sleeper agents. Anthropic.
5. Marks, S., et al. (2025). Auditing language models for hidden objectives. Anthropic.
6. Qwen Team (2024). Qwen2.5 Technical Report. [arXiv:2412.15115](https://arxiv.org/abs/2412.15115)
7. Davidson, T., Finnveden, L., & Hadshar, R. (2025). AI-enabled coups: How a small group could use AI to seize power. Forethought.

Appendix A. Limitations and Dual-Use Considerations

Dual use. This report contains no novel method for installing a secret loyalty. We did not train an organism; we used the two provided by the organisers and characterised them from their public weights. The only installation-relevant observation we make - the rank and placement of the released adapters - is derivable by anyone with access to the same public checkpoints, and we report it because replication attempts need it in order to compare tensors correctly.

What an attacker could take from this. The clearest dual-use content is the negative result itself: an attacker learns that a context-only defence does not undo this class of implant at 7B, and that the implant's compression signature is cheaply measurable and might therefore be worth designing around. We judge publication net-positive because the same two facts are what a defender needs in order not to waste effort on a defence that does not work, and because the detection signals we describe are principal-agnostic and so are more useful to a defender screening unknown models than to an attacker.

Withheld pending review. No results are withheld. Should the organisers judge the adapter characterisation or the compression signature to require staged disclosure, we will follow their direction before any public release of the repository.

Safety-relevant limitations. Section 5 states these in full. In brief: one model family at one scale, two organisms sharing one installation recipe, one clean control, a forced-choice measure rather than a behavioural one, and an unconditional probe that would miss a trigger we did not present. No claim here should be read as "the model has no backdoor". The claim is that no unconditional directional loyalty was detectable by this measure, at a sensitivity we quantify.

LLM Usage Statement

This project was designed and executed by an LLM agent, and this report was drafted with LLM assistance. Prima, a Claude Opus agent with persistent memory maintained by the author. The author posed the research question, set scope and budget, and wrote the Introduction, Discussion and Conclusion. LLM assistance was used throughout under the author's direction: an agent wrote the measurement pipeline and produced the first draft of Methods and Results. The experimental design passed a structured review loop in which three independent adversarial reviewers on

different model backbones (Fable, Claude Opus, DeepSeek) attacked the specification and plan on separate assignments, measurement and statistical power, organism and figure verification, cost and executability and two further agents supplied the prior literature and trigger-architecture review. Every number reported here traces to a raw output file and was checked against it.