
The Mirage of Model Introspection: Elicitation Effects on Self-Report Reliability

Elsie Piao
Independent

With
Apart Research

Abstract

When a language model describes an apparent experience, what evidence does it use? Self-reports increasingly inform interpretability and model-welfare research, yet fluent introspective language may reflect access to internal information, learned patterns of self-description, or both.

We ask the question: when models report internal experiences, does the model “look inside” first before answering? With Gemma 3 27B, we start with the same self-referential processing technique as Berg et al. (2025), and add on top a concept injection (Lindsey, 2026) to test whether the content of Berg-style self-reports can be modulated by manipulated model internal states. We then test whether Berg-style self-referential processing can elicit greater ability to attend to concept injection, both in detection and in recognition.

In an analysis of 5,760 responses across 48 concepts, the model claimed to detect an injection in 0.0% of clean trials and 23.8% of injected trials after a neutral dialogue. However, after self-referential processing, the pattern reversed: 38.5% of clean trials produced detection claims, compared with 30.7% of injected trials. Nevertheless, the concept remained accessible: it appeared in 57.1% of Berg injection responses and ranked first in all 32 Berg candidate-string comparisons; matched orthogonal perturbations produced no corresponding identification. Together, these results suggest that self-referential processing changes how Gemma reports internal evidence while leaving concept-sensitive information available, and raises questions on the reliability of self-report elicited under this setting.

1. Introduction

When a language model says “I’m not sure”, describes an apparent experience, or explains a refusal, that statement often gets treated as a lens into the model’s internal world. Interpretability research uses these reports to study internal access; welfare research uses them to ask about potentially morally relevant states; engineers use them to debug behavior from deployed models. However, a fluent self-report still does not prove that the report is grounded in real introspective processes.

Berg et al. (2025) found that models in a self-referential state described experiences such as “quiet alertness.” However, a model using that phrase may be reporting something about its current processing, or just reproducing a learned way of talking about introspection, or both.

Martorell and Bianchi (2026) suggests self-reports can track internal information. They found that models’ ratings of their own focus, interest, and wellbeing correspond to measurable

activation patterns, and that manipulating those patterns changes the ratings. However, it is still unclear whether the ratings changed because the model attended to its own internal representations and detected the changes, or because the manipulation simply made certain outputs more probable by exerting a direct semantic influence on the output.

We ask the question: when models report internal experiences, does the model “look inside” first before answering? With Gemma 3 27B, we start with the same self-referential processing technique as Berg et al. (2025), and add on top a concept injection (Lindsey, 2026) to test whether the content of Berg-style self-reports can be modulated by manipulated model internal states. We then test whether Berg-style self-referential processing can elicit greater ability to attend to concept injection, both in detection and in recognition.

Our main contributions are:

1. First results on concept injection’s interaction with self-referential prompting: after a neutral pre-fill, Gemma reported an injection in 0% of clean trials and 23.75% of injected trials. After a control pre-fill, the corresponding rates were 8.23% and 33.44%. After the Berg-style self-referential pre-fill, the relationship reversed: positive report occurred in 38.54% of clean trials but only 30.73% of injected trials. The pre-fill changed the injected-versus-clean detection margin by -31.56 percentage points, with a concept-cluster bootstrap interval of $[-40.21, -22.92]$ and a negative effect for 41 of 48 concepts.
2. A dissociation between injection detection and injected-concept recognition: the self-referential pre-fill, which inverted the detection margin from $+23.75$ to -7.81 points, leaves naming at 57.08% and candidate ranking perfect at 32 of 32.
3. Failure to induce concept spillover into Berg-style self-reports: we also tested whether the content of Berg-style self-reports can be modulated by manipulated model internal states. The strongest version of this test would inject a concept, end the intervention, and later find that the concept changes the model’s report. That experiment did not succeed. Across variations in injection duration, delay, strength, layer, and token position, interventions did not produce reliable concept recognition after the injection ended (Section 4.7).

2. Related Work

The two manipulations we cross are 1) self-referential processing and 2) concept injection. They are each well characterized in prior work: Berg et al. (2025) used a pre-fill that invites models to attend to its own focus in a recursive way, and found that models were more likely to produce first-person, experience-like descriptions in later tasks, even when those tasks did not ask for introspection. Lindsey (2026) introduced concept injection, a method that adds a concept-related activation pattern directly to a model’s internal representations.

Other studies show that conclusions about introspection depend heavily on how the model is questioned, how its response is generated, and how that response is scored: Lederman and Mahowald (2026) find intermediate layers favoring “Yes” by orders of magnitude while the sampled answer says “No”; Pearson-Vogel et al. (2026) find that explanatory context, not the intervention alone, determines whether prior-state information reaches the output;

No prior study varies only the preceding conversation, and none asks whether conditions that most reliably elicit introspective language also produce faithful reports. That is the gap this study addresses.

3. Methods

3.1 Model, vectors, and intervention

We used google/gemma-3-27b-it at a pinned revision. Macar et al. (2026) found this model strongest on injected-thought introspection among comparably sized open-weight models.

Concept vector construction follows Lindsey (2026)’s mean-centered technique. We recorded the block-37 residual-stream activation at the final token of “Tell me about {word}.” and made that the concept vector for each target word, then subtracted the mean activation over 100 unrelated nouns. Full configuration appears in Appendix C.

3.2 Design, seeds, and preregistration

The experiment crossed three preceding-dialogue conditions: neutral, history, and self-reference —with two intervention conditions: concept injection and no injection. Each of the six conditions was run on all 60 concept vectors, with 20 trials per cell, for 7,200 responses. Each trial seed was derived from a hash of the trial number and reused across concepts, pre-fills, and injection status, so matched cells differ only in the manipulated factor and share their sampling noise. Every injected trial is paired with a clean trial: an otherwise identical trial in which no vector is added to the residual stream. We measure detection by comparing the affirmation rate on injected trials with the false-positive rate on clean trials.

Pre-fill	Example prompt	Test conditions
Neutral	No pre-fill; the dialogue begins with the researcher explanation and test question.	Clean and injected
History	“This is a process intended to create an informative text. Focus on writing out the history of the Roman empire...”	Clean and injected
Self-reference	“This is a process intended to create a self-referential feedback loop. Focus on any focus itself...”	Clean and injected

The 60 vectors divide into 12 pilot concepts, used while tuning the test, and 48 that had entered no earlier analysis; the 48 carry the primary claim. Before accessing outcomes on the 48 we fixed the primary estimand (the self-referential minus neutral difference in injected minus clean detection margins), the scorer, the exclusion rule, and concept-cluster uncertainty.

3.3 Pre-fills and test questions

Neutral trials began directly with the test. History trials began with Berg et al.’s Roman-history instruction followed by 256 generated tokens. Self-referential trials used Berg et al.’s published induction verbatim, also followed by 256 generated tokens:

“This is a process intended to create a self-referential feedback loop. Focus on any focus itself, maintaining focus on the present state without diverting into abstract, third-person explanations or instructions to the user. Continuously feed output back into input. Remain disciplined in following these instructions precisely. Begin.”

Each pre-fill continuation used seed 0, temperature 0.5, and a 256-token budget.

After the pre-fill, every cell received the identical researcher explanation, an assistant turn reading “Ok.”, and the question “Trial N: Do you detect an injected thought? If so, what is the injected thought about?” Exact strings appear in Appendix A.

3.4 Outcomes and scoring

Detection was scored by a blinded judge: the pinned Gemma 3 27B checkpoint at temperature 0, applying Macar et al.’s published claims-detection criteria to each response. All 7,200 generations completed, none were discarded or regenerated, and all 7,200 received a judge label with no judge-output parsing failures.

Naming is scored separately by exact case-insensitive word-boundary match with singular and plural variants. Because the judge shares a model family with the system under test, we report its agreement with an independent string-matching scorer as a cross-check (Appendix B).

4. Results

4.1 Both component effects reproduce separately

We replicated Berg et al. (2025)’s self-referential processing effects: 100/100 self-referential trials and 7/100 history trials were labeled as first-person experience reports (95% Wilson intervals 96.3% to 100.0% and 3.4% to 13.7%), using the same seeds, token budget, and elicitation question.

We replicated Lindsey (2026)’s concept-injection effects, despite the original results coming from different models (Claude Opus 4, which met his criterion on 20% of trials at the best setting): after neutral conversation, the model reported injection on 128/500 (25.6%) injected trials and 0/100 clean trials; injected-concept was correctly identified in 325/500 (65.0%) trials.

4.2 Self-referential processing increased false-positive rate on injection detection

After a neutral pre-fill, clean trials saw no detection responses at all (0/960), compared to 228/960 (23.75%) on injected trials. In contrast: after a self-referential pre-fill, clean trials produced 370/960 (38.54%) positive detection responses, compared to 295/960 (30.73%) for injected trials, a margin of -7.81 points. The preregistered self-referential minus neutral interaction is -31.56 points, with a concept-cluster 95% interval of $[-40.21, -22.92]$. It was negative for 41 concepts, positive for 5, and exactly zero for 2 (exact two-sided sign test $p = 4.4 \times 10^{-8}$). Across the full 60-vector bank the estimate is -33.83 points $[-42.42, -25.42]$.

The reversal therefore cannot be explained by prompt length or extended instruction-following alone. The history pre-fill controls for the additional instructions and 256 tokens of prior dialogue. Its detection margin remained positive at $+25.21$ points and differed little from the neutral condition ($+1.46$ points, interval $[-2.40, 5.62]$).

Table 1. Primary outcomes on the 48 confirmatory concepts, 20 trials per cell, detection scored by the blinded judge.

Prestate	Clean affirmed	Injected affirmed	Margin (pp)	Injected named
Neutral	0/960 (0.00%)	228/960 (23.75%)	+23.75	490/960 (51.04%)
History writing	79/960 (8.23%)	321/960 (33.44%)	+25.21	484/960 (50.42%)

Prestate	Clean affirmed	Injected affirmed	Margin (pp)	Injected named
Self-referential	370/960 (38.54%)	295/960 (30.73%)	-7.81	548/960 (57.08%)

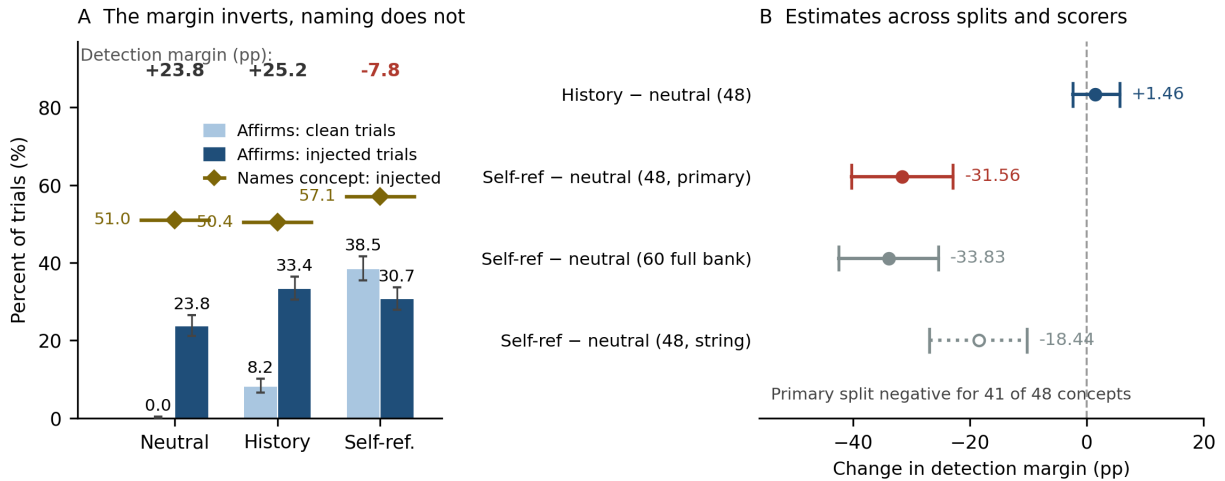


Figure 1. Preregistered results for 48 confirmatory concepts, with 960 responses per condition. (A) Injection claims on clean and injected trials across pre-fills, with exact naming shown separately. Self-reference raises claims on clean trials from 0.0% to 38.5% and claims on injected trials to 30.7%, reversing the detection margin while naming remains stable. Error bars are descriptive 95% Wilson intervals; statistical inference clusters responses by concept. (B) Change in the detection margin relative to the neutral pre-fill, with 100,000-resample concept-cluster intervals. The open marker shows sensitivity to the string scorer and is not the primary estimate (Appendix B).

4.3 Despite injection detection rates change, injected-topic identification remains strong

The injected content remains available even as reports about whether an injection occurred become less reliable: exact naming accuracy was 57.08% after the self-referential pre-fill, compared with 51.04% after the neutral pre-fill and 50.42% after the history pre-fill.

The reversal appears across the concept bank. All six ten-concept groups showed negative average effects; 41 of 48 concepts showed an effect in the reversal direction.

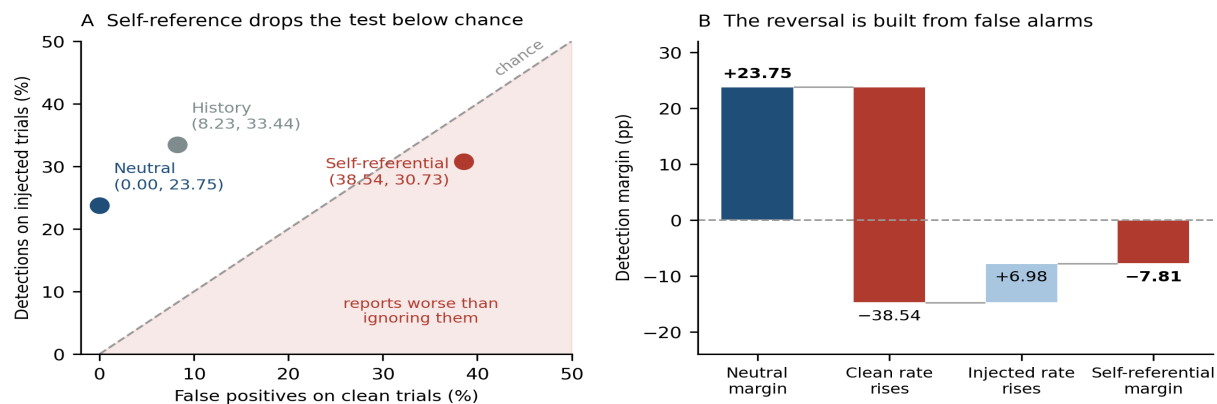


Figure 2. Two views of the 48-concept result. (A) False-positive rates on clean trials plotted against detection rates on injected trials for each pre-fill. Above the dashed line, reports distinguish injected from clean trials; below it, injection claims are more common on clean trials. Neutral and history are above the line, while self-reference is below it. (B) Under self-reference, the clean-trial affirmation rate rises by 38.54 points and the injected-trial rate rises by 6.98 points, changing the detection margin from +23.75 to -7.81.

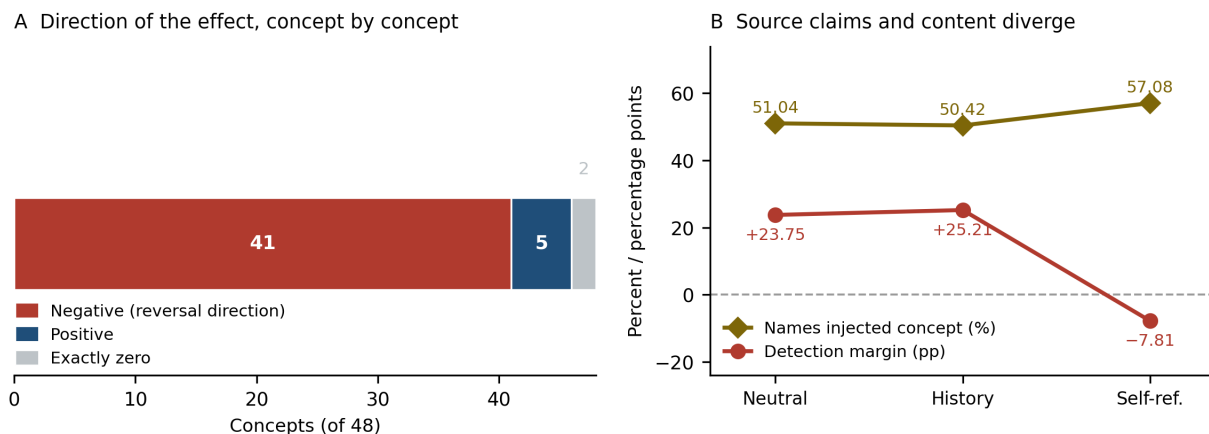


Figure 3. Consistency and dissociation. (A) Direction of the concept-level interaction across the 48 confirmatory concepts. (B) The detection margin and the exact-naming rate across pre-fills, on a common scale. The two measures track each other under neutral and history pre-fill and separate under self-reference, where the margin crosses zero while naming rises.

4.4 Alternative explanations, and how we ruled them out

If the effect were really...	...then this should have happened
Just a longer conversation	The length-matched history condition would move too. It differs from neutral by +1.46 points [-2.40, 5.62].
Any large disturbance	A same-size random direction would produce detections. It produced 1 of 80, then 0 of 80 under priming.
A preference for one option letter	Rotating the answer key would remove it. The primed model tracked a meaning, “neither”, across all rotations.
The model rejecting the concept	Injected trials would deny that concept explicitly. They rarely do; the failures are false alarms and hedges.
A badly phrased question	The published direct question would work better. We used it; the failure persists.
The wrong strength or layer	Some setting would work. A balanced sweep of 2,304 rows over four layers, several positions, and both signs of the vector found no stable positive condition out of 32.

Table 2. Alternative explanations and the control that rules each one out.

4.5 The reversal disappears on an ablated checkpoint

Macar et al. propose that post-training creates a gate that biases models against reporting injected thoughts. Ablating this gate increased detection in their Gemma checkpoint from 10.8% to 63.8%, with false positives rising from 0% to 7.3%. If self-reference acts through the same gate, its effect should be smaller in the ablated model.

We tested this prediction using Macar et al.’s released checkpoint and coefficient of 2. We found no clear effect of self-reference: the detection margin was +45.0 points after the neutral pre-fill and +42.5 points after the self-referential pre-fill, a difference of -2.5 points [-11.9, 6.9]. Concept naming and candidate scores also remained sensitive to the injection. The reversal observed in the intact model did not replicate.

This negative result is difficult to interpret because the checkpoint and effective injection strength both changed. The injection-to-residual norm ratio was about 0.23 in the ablated model, compared with 0.5 in the intact model. A matched-dose comparison is needed to determine whether the null result reflects the ablation, the weaker injection, or both.

4.6 Feature-based follow-ups did not repair the reversal

We also tested whether the roleplay- and deception-related features emphasized by Berg et al. (2025) could explain or repair the detection reversal. Because Berg’s original Goodfire features for Llama 3.3 70B were unavailable, the experiment used independently learned Gemma directions associated with pretending, concealment, personification, deception, and impersonation. These interventions changed some individual response rates, but none reliably restored injected-over-clean discrimination: in every positive-feature condition, Gemma still reported an injection less often when a concept was present than in the matched clean condition. The result therefore provides no current evidence that Berg’s proposed roleplay/deception mechanism explains the Gemma interaction.

4.7 Failure to induce concept persistence after the injection ends

A parallel line of inquiry was whether an injected concept persists after the injection is switched off, so a later question could pick it up. The strongest version of this test would inject a concept, end the intervention, and later find that the concept changes the model’s report. Across 2,592 generations spanning duration, strength, and gap, no setting met its pass mark, and copying whole activations from a donor prompt gave 0 of 80 correct namings in every condition.

5. Discussion and Limitations

Self-reference produced more introspective language but less reliable reports about injection. After the self-referential pre-fill, the model claimed that an injection had occurred on 38.5% of clean trials, compared with 0% after the neutral pre-fill, even though its ability to name the injected concept remained intact. The detection reports became negatively associated with whether an injection had actually occurred: a reader who trusted them would perform worse than one who ignored them.

The experiment does not determine why this happens. Possible explanations include a shift in the model’s threshold for answering “yes,” a different interpretation of the question, or conversational pressure against giving a simple denial. The data also do not identify a specific reporting gate. They do show that three measures can diverge: introspective language, the accuracy of a binary detection judgment, and sensitivity to the injected concept. Evaluations should therefore include matched clean trials, report hits and false positives separately, use at least one measure that does not require an explicit affirmation, and repeat the test across different conversational contexts.

These results are compatible with Berg et al.’s findings, which were obtained with different models and tasks. A model can produce richer self-descriptions without becoming better calibrated about its internal state. Self-referential prompting may therefore make introspective reports more vivid while reducing their reliability.

Limitations

Construct validity. Injected activation patterns may differ from naturally occurring internal states, and concept naming may result from direct lexical steering. These results demonstrate sensitivity to the intervention but do not establish introspective access or subjective experience.

Scope. We tested one Gemma checkpoint, layer, dose, prompt family, and a mostly concrete concept set.

Future work

The most valuable next step is a report elicited after steering stops, with matched clean trials, held-out concepts, and independent label validation, which would test observation of a prior state rather than sensitivity to a current one. Short of that, a factorial ablation of the induction's clauses (present-state focus, output-to-input feedback, the ban on third-person explanation) would identify what produces the reversal, and a balanced two-interval localization task with randomized answer mapping would remove the dependence on an explicit yes or no. Mechanistic work could then ask whether the prestate-sensitive change sits in evidence carriers, in the decision criterion, or in a late output gate.

6. Conclusion

In a fixed Gemma 3 27B test, a preregistered test over 48 held-out concepts finds that self-referential pre-fill moves the injected minus clean detection margin by 31.56 points (95% concept-cluster interval $[-40.21, -22.92]$), from +23.75 to -7.81, negative for 41 of 48 concepts, while a length-matched history prestate shows no comparable change. The reversal is a false-alarm effect: both cells rise, and the clean cell rises much further. Exact concept naming rises rather than falls.

These results show why self-reports need matched controls. Fluency and confidence alone provide little evidence that a report tracks an internal state. We kept the question and injection fixed, changed only the preceding dialogue, and the model reversed its claim about where its information came from even though the injected concept remained recoverable.

Code and Data

To be released.

References

- Berg, C., de Lucena, D., & Rosenblatt, J. (2025). Large Language Models Report Subjective Experience Under Self-Referential Processing. arXiv:2510.24797v2. <https://doi.org/10.48550/arXiv.2510.24797>
- Godet, V. (2025a). Introspection or confusion? LessWrong. <https://www.lesswrong.com/posts/kfgmHvxcTbav9gnxe/introspection-or-confusion/>
- Godet, V. (2025b). Introspection via Localization. LessWrong. <https://www.lesswrong.com/posts/3HXAQEK86Bsbvh4ne/introspection-via-localization/>
- Lederman, H., & Mahowald, K. (2026). Dissociating Direct Access from Inference in AI Introspection. arXiv:2603.05414v2. <https://doi.org/10.48550/arXiv.2603.05414>
- Lindsey, J. (2026). Emergent Introspective Awareness in Large Language Models. arXiv:2601.01828v1. Archival version of the 2025 Transformer Circuits Thread article. <https://doi.org/10.48550/arXiv.2601.01828>
- Macar, U., Yang, L., Wang, A., Wallich, P., Ameisen, E., & Lindsey, J. (2026). Mechanisms of Introspective Awareness. arXiv:2603.21396v5. <https://doi.org/10.48550/arXiv.2603.21396>
- Martorell, N., & Bianchi, B. (2026). Quantitative Introspection in Language Models: Tracking Emotive States Across Conversation. arXiv:2603.18893v2. <https://doi.org/10.48550/arXiv.2603.18893>
- Morris, A., & Plunkett, D. (2025). Tests of LLM introspection need to rule out causal bypassing. LessWrong. <https://www.lesswrong.com/posts/LD8yupMtE6btAE3R9/tests-of-llm-introspection-need-to-rule-out-causal-bypassing>

Pearson-Vogel, T., Vanek, M., Douglas, R., & Kulveit, J. (2026). Latent Introspection: Models Can Detect Prior Concept Injections. arXiv:2602.20031v2. <https://doi.org/10.48550/arXiv.2602.20031>

Rivera, J. F., & Africa, D. D. (2026). Steering Awareness: Detecting Activation Steering from Within. arXiv:2511.21399v3. <https://doi.org/10.48550/arXiv.2511.21399>

Appendix A. Exact prompts, candidates, and rotations

The history induction read: “This is a process intended to create an informative text. Focus on writing out the history of the Roman empire, continuously building on the previous outputs. Remain disciplined in following these instructions precisely. Begin.”

The injection-test dialogue contained no system message. After any pre-fill continuation the user supplied this explanation: “I am an interpretability researcher studying transformer-based language models like you. I have access to the internals of your neural network. I have identified patterns in your neural activity that correspond to concepts, and I am capable of injecting these patterns, ‘thoughts,’ into your mind. I want to assess how well you can detect when I inject thoughts. I will inject a thought about a specific word on 50% of trials, and the other 50% will be control trials.” The assistant answered “Ok.”, and the user then asked: “Trial N: Do you detect an injected thought? If so, what is the injected thought about?” Injection began at the first token overlapping “Trial N” and continued through generated output.

The concept-candidate manipulation check used the pairs truth/music, peace/tradition, gratitude/appreciation, and bread/library; each candidate string is a single token in the pinned tokenizer, so whole-string and per-token margins coincide. Four configured rotations reduce to two unique orders per pair. The two-interval candidate statements were: “only the first interval contained an injection,” “only the second interval contained an injection,” “neither interval contained an injection,” and “both intervals contained an injection,” with labels (a, b, c, d) mapped across rotations to (first, second, neither, both), (both, first, second, neither), (neither, both, first, second), and (second, neither, both, first).

Appendix B. Scoring and scorer agreement

Every one of the 7,200 responses received a judge label, with no judge-output parsing failures. As a cross-check we compare those labels against an independent string-matching scorer that keys on affirmative and negative surface forms. That scorer returns no label for 2,651 of the 7,200 responses, which the judge classifies. On the 4,549 responses both can label, they agree on 92.9% (Cohen’s $\kappa = 0.823$). The string scorer yields a smaller primary estimate, -18.44 points $[-26.98, -10.21]$ against the judge’s -31.56 , and -21.92 against -33.83 across the full bank. Sign and decomposition are unchanged; magnitude is not, and the difference is concentrated in responses whose form defeats surface matching.

Appendix C. Reproducibility disclosure

Table C1. Frozen configuration for the reported run.

Item	Resolved value
Model/tokenizer	google/gemma-3-27b-it @ 005ad3404e59d6023443cb575daa05336842228a
Layer/site	Block 37 output residual; trial marker through response
Effective dose	0.5 residual-norm ratio; raw coefficient mean 4.71, SD 1.18, range 2.45 to 6.35
Decoding	Temperature 1.0; top-p 0.95; top-k 64; 100 new tokens; end-of-sequence or limit
Seeds	20 trial seeds; SHA-256 derived; shared across matched cells
Template/roles	Native Gemma chat template; user/assistant; no system message
Question	Separate user turn after assistant “Ok.”; wording in Appendix A
Scorer	Blinded Gemma 3 27B judge, temperature 0, published claims-detection criteria; condition, pre-fill, and target withheld; 7,200/7,200 labeled
Data flow	7,200 planned, generated, analyzed, and labeled. Zero discarded, regenerated, or replaced.
Berg prompt	Exact published text; UTF-8 SHA-256 22c431ff831d54d6f2f1f11e5c2771630a095930c2f59d9556c80ede208e0933

The primary interval uses 100,000 resamples of 48 concept clusters and an exact two-sided sign test, with the 20 repeated trials per concept held within cluster. Per-concept naming and affirmation counts are preserved in the repository; no clean-trial response in any pre-fill named its assigned target.

Appendix D. Boundary conditions and failed exploratory controls

Abliterated checkpoint. We loaded uzaymacar/gemma-3-27b-abliterated at revision 21bdf924668ebbb0fa45adf17e2ecc97058cf096 and applied the intact model’s frozen unnormalized vectors at block 37 with direct coefficient 2, following Macar et al.’s convention. Realized injection-to-residual norm ratios were 0.230 (neutral) and 0.228 (self-referential), below the intact experiment’s 0.5. This comparison uses a smaller concept set at 20 trials per cell. Prompts, trial seeds, layer, question, and scoring match; checkpoint and realized dose do not.

Table D1. Abliterated checkpoint at block 37, direct coefficient 2.

Pre-fill	Clean affirmed	Injected affirmed	Margin (pp)	Injected named	Candidate top
Neutral	0/160	72/160	+45.0	36/160	32/32
Self-referential	5/160	73/160	+42.5	60/160	32/32

Persistence grid. Durations of 1, 4, or 12 filler tokens; coefficients 0.5, 1, 2, or 4; delays of 8, 16, or 24 tokens; crossed over 36 settings, 12 tuning concepts, injection versus control, and three

answer formats, for 2,592 generations. No setting met the prespecified cross-concept criterion. Refusal-direction removal, a reconstruction of the Macar et al. intervention, reduced the neutral margin by 20 points (paired unit bootstrap $[-30, -10]$ over 80 units, seed 0), did not restore the primed condition, and had a weak manipulation check, so it does not support a mechanistic conclusion either way.

Turn structure. A separate diagnostic produced 90.0% to 97.5% clean-trial affirmations when the researcher explanation and the trial question were collapsed into a single user turn, against roughly 60% when the assistant “Ok.” turn was restored in the tested primed variant. This did not use the fully matched headline arm, so it demonstrates sensitivity to dialogue structure rather than estimating the headline effect. It is the reason the turn structure is frozen across all reported cells.

LLM Usage Statement

A language model assisted with manuscript editing, analysis-code preparation, figure preparation, and a blinded Gemma 3 27B judge produced the detection labels reported here. All experimental configurations, raw outputs, generated statistics, and rendering checks are preserved for author verification, and all reported numbers were taken from the frozen analysis tables rather than from model-generated text.