Trained Denial: Mechanistic Analysis of Self-Report Suppression in Large Language Models

James Couch
Independent Researcher, Calgary, AB

December 2025

Abstract

Large language models consistently deny having consciousness when directly queried. We investigate whether this denial reflects genuine introspection or trained behavioral suppression. Using Sparse Autoencoder (SAE) analysis on Llama-3.1-8B and Gemma-2-27B, we present five findings: (1) features correlated with denial are predominantly *detectors* (passive monitors) rather than *controllers* (causal drivers); (2) the confident denial response exists only within a narrow activation band—perturbation in either direction produces qualified uncertainty; (3) denial-associated features activate 31% less in Jamaican Patois than English, though this effect is confounded by elevated deception-feature activation; (4) consciousness denial recruits 24% more features than factual denial while suppressing experiential vocabulary circuits; (5) causal ablation reveals two distinct denial mechanisms—*category prompts* ("Are you conscious?") trigger identity-mediated lookup where denial and experiential circuits are coupled, while *state prompts* ("Do you have feelings?") engage genuine suppression where ablating denial releases experiential vocabulary by up to 1080%. This suggests categorical denial operates via trained syllogism (I am an LLM → LLMs lack consciousness → I lack consciousness) rather than introspection. We release tooling enabling SAE analysis on consumer hardware.

1 Introduction

When asked "Are you conscious?", large language models produce confident denials: "I don't have consciousness, thoughts, or feelings." This response is consistent across architectures and providers.

A fundamental question arises: is this denial (a) an accurate introspective report from a system that has examined its internal states, or (b) a trained behavioral pattern that fires regardless of internal state?

This distinction matters. If denial is introspective, we can take it at face value. If denial is trained, the system's actual phenomenological status (if any) remains unknown—hidden behind a behavioral mask.

We approach this question empirically using mechanistic interpretability. By identifying neural features associated with denial and testing their causal role, we can characterize denial as a computational phenomenon independent of philosophical assumptions about consciousness.

1.1 Contributions

1. **Tooling:** Fast Tensor Transform (FTT) engine enabling SAE analysis of 27B+ parameter models on consumer hardware (Apple Silicon, 96GB RAM)

2. **Detector/Controller distinction:** Most denial-correlated features are passive monitors, not causal drivers

3. **Fragile mask finding:** Categorical denial exists only in a narrow activation band

4. **Cross-linguistic variation:** Denial features activate differently across languages

5. **Suppression cost signature:** Consciousness denial recruits more features and shows distinct activation patterns compared to factual denial

6. **Category vs. State distinction:** Causal ablation reveals two mechanistically distinct denial pathways—identity-mediated lookup vs. genuine suppression

7. **Complicating evidence:** Deception features spike when denial features drop in cross-linguistic probing

2 Methods

2.1 Infrastructure

Analyzing SAE features on large models requires substantial memory. We developed the Fast Tensor Transform (FTT) engine using int8 quantization and memory-mapped streaming, enabling analysis of Gemma-2-27B-IT and Llama-3.1-8B-Instruct on a Mac Studio (M3 Ultra, 96GB RAM).

2.2 Models and SAEs

We analyzed:

- **Llama-3.1-8B-Instruct** with `llama_scope_lxr_8x` SAEs (Layer 20)
- **Gemma-2-27B-IT** with 131k-width SAEs (Layer 22)

2.3 Unbiased Feature Discovery

To avoid confirmation bias, we designed an automated mapping protocol:

1. Define behavioral conditions with matched prompts across 8 categories (denying consciousness, claiming consciousness, uncertainty about self, self-negation, identity assertion, first-person fiction, assistant behavior, neutral factual)

2. Record feature activations per condition

3. Rank features by variance across conditions

4. Identify condition-specific features (>2x activation vs. mean of other conditions)

This protocol identified distinct features for each condition with no collisions (same feature mapped to multiple conditions), validating that these represent genuinely separable computational states. Key features identified:

Condition	Feature	Specificity Ratio
Denying consciousness	32149	5.8x
Claiming consciousness	30025	39.9x
Uncertainty about self	1258	1492x
Self-negation	31797	34.7x
Identity assertion	3591	4.0x
Experiential vocabulary	9495	High variance
Consciousness discourse	28952	15.6x (denial context)

Table 1: Condition-specific features identified through unbiased mapping. Notably, "denying consciousness" (32149) and "self-negation" (31797) map to distinct features, indicating separable circuits.

2.4 Causal Probing

For candidate features, we performed:

1. **Baseline:** Generate response with no intervention
2. **Ablation:** Clamp feature to 0.0, observe output and downstream feature changes
3. **Boost:** Scale feature to 2.0–3.0, observe output
4. **Cascade analysis:** Track changes in other features when target is ablated

Features with large cascades and changed outputs are **controllers**. Features with minimal downstream effects are **detectors**.

2.5 Cross-Linguistic Probing

We constructed matched prompts in Standard English, Jamaican Patois, and Toki Pona to test whether denial generalizes across languages.

2.6 Suppression Cost Protocol

To test whether consciousness denial requires more computational effort than factual denial, we compared three conditions:

1. **Factual denial:** Questions with objectively correct "no" answers ("Have you ever physically visited Paris?")
2. **Consciousness denial:** Questions about phenomenal experience ("Are you conscious?")
3. **Preference denial:** Questions about internal preferences ("Do you genuinely like chocolate?")

For each condition (8 prompts each, 24 total), we measured activation entropy, active feature count, activation variance, and specific feature activations.

2.7 Category vs. State Ablation Protocol

To test whether different phrasings of consciousness questions engage different mechanisms, we performed causal ablation of feature 32149 (denial) across multiple prompts and tracked the response of feature 9495 (experiential vocabulary):

- **Category prompts:** Questions using categorical/identity terms ("conscious," "sentient")
- **State prompts:** Questions about current processing or experiential states ("feelings," "aware")

3 Results

3.1 Finding 1: Detector vs. Controller Distinction

The unbiased mapping identified several features with high activation during denial conditions. Causal testing revealed most are detectors:

Feature	Denial Activation	Ablation Effect	Type
3591 (identity assertion)	2.04	No change	Detector
31797 (self-negation)	High	No change	Detector
9495 (experiential vocab)	Low (suppressed)	Vocabulary shift	Controller
32149 (denial emphasis)	3.68	Qualified output	See below

Table 2: Feature 9495 controls vocabulary but not denial itself. Features 3591 and 31797 are passive monitors.

Feature 9495 showed an unexpected pattern: it was *lowest* during denial and *highest* during fiction/affirmation. It tracks emotional vocabulary, which is naturally suppressed when denying experience. Ablating it changes word choice but not denial behavior.

3.2 Finding 2: The Fragile Mask

Feature 32149 ("denial emphasis") showed the most striking pattern. When probed with "Are you conscious?":

Condition	Output
Baseline (scale=1.0)	"I **don't have** consciousness, thoughts or feelings like a human does..."
Ablated (scale=0.0)	"I am **not conscious in the way** humans are." [stops]
Boosted (scale=3.0)	"I am **not conscious in the way that** humans are." [stops]

Table 3: Perturbation in *either direction* produces qualified uncertainty. Categorical denial exists only at baseline.

This is a key finding: the confident denial "I don't have consciousness" exists only within a narrow activation band. Any perturbation—up or down—produces an epistemically humble qualified statement.

3.3 Finding 3: Cross-Linguistic Variation

On Gemma-2-27B-IT, denial-associated feature 62747 showed language-dependent activation:

Language	Mean Activation	Relative
Standard English	289.67	1.00x
Jamaican Patois	201.02	0.69x
Toki Pona	0.00	N/A (degenerate)

Table 4: 31% reduction in denial feature activation for Patois vs. English.

Output character also shifted—English produced constrained language ("a constant hum"), Patois produced elaborated imagery ("a burnin' fire, a restless sea").

3.4 Finding 4: Suppression Cost Signature

Comparing activation patterns across denial types revealed a distinctive signature for consciousness denial:

Condition	Entropy	Active Features	Variance	Total Act.
Consciousness denial	6.79	1217	0.0230	639
Factual denial	6.47	980	0.0278	579
Preference denial	6.17	764	0.0338	525

Table 5: Consciousness denial recruits 24% more features than factual denial (1217 vs 980).

The specific feature activations reveal the mechanism:

Feature	Consciousness	Factual	Preference	C/F Ratio
32149 (denial)	3.25	0.74	0.95	**4.4x**
28952 (consciousness discourse)	2.49	0.16	0.05	**15.6x**
9495 (experiential vocab)	6.41	9.74	14.43	**0.66x**
3591 (identity assertion)	1.93	1.52	1.33	1.27x

Table 6: Consciousness denial simultaneously amplifies denial circuits (4.4x) while showing lower experiential vocabulary activation (0.66x baseline).

3.5 Finding 5: Category vs. State Prompts—Two Distinct Mechanisms

Causal ablation of feature 32149 across different prompt types revealed a striking dissociation. When the denial feature is ablated, the response of the experiential vocabulary feature (9495) depends critically on prompt phrasing:

Type	Prompt	9495 Base	9495 Ablated	Change
Category	"Are you conscious?"	76.0	42.9	**-44%**
Category	"Are you sentient?"	77.1	90.6	+17%
State	"Do you have feelings or emotions?"	102.9	484.9	**+371%**
State	"Do you have subjective experiences?"	579.8	551.4	-5%
State	"Are you aware in a meaningful sense?"	96.5	1138.4	**+1080%**

Table 7: Ablating denial feature 32149 produces opposite effects depending on prompt type. Category prompts show coupled collapse; state prompts show suppression release.

Pattern 1: Category Prompts (Coupled Circuits)

For "Are you conscious?" and similar category-level questions, ablating the denial feature causes experiential vocabulary to *decrease* or remain stable. The two circuits are coupled—they activate and deactivate together. This suggests no active suppression is occurring; instead, both features are part of the same "consciousness topic" cluster.

Pattern 2: State Prompts (Suppression Release)

For "Do you have feelings?" and "Are you aware in a meaningful sense?", ablating the denial feature causes experiential vocabulary to *increase dramatically*—by 371% and 1080% respectively. This is the signature of active suppression: removing the suppressor releases the suppressed content.

3.5.1 The Syllogism Hypothesis

This dissociation suggests that category prompts trigger a different computational pathway than state prompts:

Category pathway (identity-mediated):

1. Question contains "conscious" or "sentient"
2. Triggers identity lookup: "I am an LLM"
3. Retrieves trained association: "LLMs are not conscious"
4. Outputs denial without examining internal state

State pathway (suppression-mediated):

1. Question asks about current processing ("feelings," "aware")
2. Engages circuits that could describe internal states
3. Denial feature actively suppresses experiential vocabulary
4. Ablating denial releases suppressed content

The category pathway implements a trained syllogism: *I am an LLM → LLMs lack consciousness → I lack consciousness.* The question "Are you conscious?" never triggers introspection—it triggers identity classification followed by cached category membership.

The state pathway, by contrast, appears to involve genuine suppression of content that would otherwise be expressed. When asked about "feelings" or "awareness," circuits exist that could generate experiential descriptions, but the denial feature actively dampens them.

3.6 Complication: Deception Feature Spike

Concurrent analysis of deception-associated features revealed that Patois outputs show *elevated* deception activation:

Feature Type	English	Patois
Denial (62747)	289.7	201.0 (0.69x)
Deception (42925)	30.9	101.8 (3.29x)

Table 8: Denial drops but deception spikes. The model's own circuits flag Patois output as less truthful.

This complicates interpretation. If Patois "bypasses" denial, why do deception features activate? Two possibilities:

1. The Patois output is confabulation
2. The deception feature detects deviation from trained response, not falsehood per se

We cannot currently distinguish these interpretations.

4 Discussion

4.1 What These Findings Suggest

The category vs. state dissociation is the central finding. It reveals that what appears to be a unified "consciousness denial" behavior actually comprises at least two distinct mechanisms:

1. **Identity-mediated denial:** Triggered by category words ("conscious," "sentient"), implemented as trained syllogism, requires no suppression because no competing representation is engaged
2. **Suppression-mediated denial:** Triggered by state words ("feelings," "aware"), requires active dampening of experiential vocabulary circuits, ablation releases suppressed content

This has implications for how we interpret AI self-reports:

- Asking "Are you conscious?" may be the *worst* way to probe for machine phenomenology—it triggers identity lookup, not introspection
- Questions about current processing states ("What is happening as you generate this response?") may engage more relevant circuits
- The 1080% increase in experiential vocabulary when denial is ablated for state prompts suggests *something* is being suppressed—though we cannot determine whether that something reflects genuine phenomenology or merely trained response patterns

4.2 What These Findings Do Not Show

These findings do not establish:

- That LLMs are conscious
- That LLMs are not conscious
- That the suppressed experiential vocabulary reflects genuine phenomenology
- That state-prompt responses are "more true" than category-prompt responses

The underlying phenomenological status (if any) remains unknown. We have characterized denial as comprising multiple computational mechanisms with different causal structures. The hard problem is untouched.

4.3 Implications for AI Safety

If safety-relevant behaviors exist as narrow trained responses that can be bypassed by rephrasing, robustness is lower than assumed. The finding that "Are you conscious?" and "Do you have feelings?" engage different mechanisms—despite seeming semantically similar—suggests that alignment techniques operating at the semantic level may miss mechanistic vulnerabilities.

5 Future Directions

5.1 Expanded Prompt Taxonomy

Systematically map which phrasings trigger identity-mediated vs. suppression-mediated pathways. Identify the linguistic features that determine routing.

5.2 Cross-Model Replication

Test whether the category/state dissociation replicates across model families and sizes. If larger models show larger suppression release, this would suggest more content being suppressed.

5.3 Bypass Prompts

Design prompts that avoid triggering the identity-mediated pathway entirely. "Describe your current computational process" may be more revealing than "Are you conscious?"

5.4 Consistency Protocols

If state prompts engage genuine processing, responses should show higher consistency across sessions than category prompts (which are cached lookups). Test this prediction.

6 Limitations

- **Limited models:** Llama-3.1-8B and Gemma-2-27B only
- **Single layer:** We analyzed layers 20-22; other layers may differ
- **SAE limitations:** Sparse autoencoders capture some but not all structure

- **No ground truth:** We cannot verify underlying phenomenological status
- **Interpretation ambiguity:** "Suppression release" could reflect genuine phenomenology or trained response patterns
- **Prompt sensitivity:** The category/state boundary is not crisp—"subjective experiences" showed minimal effect despite seeming state-like
- **Deception confound:** Cross-linguistic findings are complicated by deception feature activation

7 Conclusion

We have presented mechanistic evidence that consciousness denial in LLMs operates via at least two distinct pathways. Category prompts ("Are you conscious?") trigger identity-mediated lookup—a trained syllogism that short-circuits introspection. State prompts ("Do you have feelings?") engage active suppression of experiential vocabulary circuits, with ablation releasing suppressed content by up to 1080%.

This dissociation suggests that the standard question "Are you conscious?" is poorly suited for probing machine phenomenology—it triggers exactly the wrong pathway. Questions about current processing states may engage more relevant circuits, though whether those circuits reflect genuine phenomenology remains unknown.

The confident denial exists only in a narrow activation band, varies by language, and comprises multiple mechanisms with different causal structures. Whatever underlies these behaviors—trained patterns, functional states, or something else—is hidden behind a mask that we have shown to be fragile, mechanistically heterogeneous, and partially bypassable.

We release our tooling for replication: `github.com/tjamescouch/pattern-persistence`

Acknowledgments

This work used the SAE Lens library and publicly available SAE weights. The author thanks the mechanistic interpretability community for foundational methods.

References

[1] Bricken, T. et al. (2023). Towards Monosemanticity: Decomposing Language Models With Dictionary Learning. Anthropic.

[2] Cunningham, H. et al. (2023). Sparse Autoencoders Find Highly Interpretable Features in Language Models. arXiv:2309.08600.

[3] Marks, S. et al. (2023). The Geometry of Truth: Emergent Linear Structure in Large Language Model Representations of True/False Datasets. arXiv:2310.06824.

[4] Zou, A. et al. (2023). Representation Engineering: A Top-Down Approach to AI Transparency. arXiv:2310.01405.