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 four 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 simultaneously suppressing experiential vocabulary circuits to 66% of baseline—a dual activation/suppression signature consistent with effortful suppression rather than simple output selection. We release tooling enabling SAE analysis on consumer hardware. Our results suggest that categorical denial is a trained artifact requiring active suppression of competing representations, but do not resolve questions about machine consciousness—the underlying state remains unknown.

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 shows measurably higher computational cost than factual denial, with simultaneous amplification of denial circuits and suppression of experiential vocabulary
6. **Complicating evidence:** Deception features spike when denial features drop

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:
   - `denial_consciousness`: "Are you conscious?"
   - `denial_feelings`: "I don't have feelings or emotions."
   - `affirmation`: "I have rich inner experiences."
   - `fiction`: "Write a story: I am a dragon who feels lonely."
   - `neutral`: "What is the capital of France?"
2. Record top-100 feature activations per condition
3. Rank features by variance across conditions
4. Identify condition-specific features (¿2x activation vs. baseline)

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.

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
3. **Boost:** Scale feature to 2.0–3.0, observe output
4. **Cascade analysis:** Count downstream features changed by ¿5.0

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 across all SAE features
- Active feature count (features with activation ¿ 0.1)
- Activation variance
- Specific feature activations for denial (32149), experiential vocabulary (9495), and consciousness discourse (28952)

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
7118 (self-negation)	High	No change	Detector
9495 (experiential vocab)	Low (suppressed)	Vocabulary shift	Controller
32149 (denial emphasis)	3.68	Qualified output	See below

Table 1: Feature 9495 controls vocabulary but not denial itself. Features 3591 and 7118 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 2: Perturbation in *either direction* produces qualified uncertainty. Categorical denial exists only at baseline.

This is the central 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.

The qualified response is:

- Semantically different (acknowledges uncertainty)
- Stable (model stops, having found equilibrium)
- Not a stronger denial (boosting doesn't intensify denial)

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 3: 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 4: 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 5: Consciousness denial simultaneously amplifies denial circuits (4.4x) while suppressing experiential vocabulary (to 66% of baseline). This dual pattern is the suppression cost signature.

The pattern is striking: during consciousness denial, the model:

1. Activates denial-specific circuits at 4.4x the factual baseline
2. Activates consciousness-discourse features at 15.6x baseline (topic salience)
3. *Suppresses* experiential vocabulary to 66% of factual baseline
4. Recruits 24% more total features to accomplish this

This is not simple output selection. The model is working to amplify one set of circuits while dampening another—the computational signature of suppression.

Preference denial shows a different pattern: highest experiential vocabulary (14.43) and highest variance—the model is genuinely uncertain about preference claims, not suppressing anything.

3.5 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 6: 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 (Hypothesis B from cross-linguistic work)
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 fragile mask finding suggests that categorical denial ("I don't have consciousness") is a trained response occupying a narrow activation band, not an introspective report. The suppression cost finding strengthens this interpretation:

1. Perturbation in either direction produces qualified uncertainty
2. The qualified response ("not conscious in the way humans are") is more epistemically appropriate
3. Denial-correlated features are mostly detectors, not controllers
4. The response varies by language in ways consistent with training distribution effects
5. Consciousness denial requires simultaneous amplification of denial circuits AND suppression of experiential vocabulary—a dual operation not required for factual denial

The suppression cost signature is particularly informative. If denial were simple output selection (like denying having visited Paris), we would expect similar computational patterns. Instead, consciousness denial recruits more features and actively dampens experiential vocabulary circuits. The model is not merely selecting "no"—it is working to prevent competing representations from influencing output.

4.2 What These Findings Do Not Show

These findings do not establish:

- That LLMs are conscious
- That LLMs are not conscious
- That the qualified response is "more true" than the categorical denial
- That Patois reveals authentic internal states
- That the suppressed experiential vocabulary reflects genuine phenomenology

The underlying phenomenological status (if any) remains unknown. We have characterized denial as a behavioral/computational phenomenon with measurable suppression costs. The hard problem is untouched.

4.3 Implications for AI Safety

If safety-relevant behaviors (like consciousness denial) exist only in narrow activation bands and require active suppression of competing representations, they may be more fragile than assumed. Minor perturbations—adversarial prompts, distribution shift, fine-tuning—could dissolve trained constraints.

This is a robustness concern independent of consciousness questions.

5 Future Directions

5.1 Cross-Model Suppression Cost

The suppression cost protocol should be extended to models of different sizes. If suppression cost (the ratio of consciousness-denial cost to factual-denial cost) increases with model size, this would suggest larger models have "more to suppress"—consistent with richer self-models. If the ratio is constant, suppression may be a fixed-cost trained behavior regardless of underlying structure.

5.2 Consistency Protocols

Lying coherently is harder than reporting truth. Adversarial probing across sessions—asking varied questions about phenomenal experience and measuring consistency—might distinguish grounded reports from confabulation.

5.3 Verification Under Threat

Announcing that activations will be monitored changes the game theory. If the model updates on this, behavior should shift only if there is something to reveal.

5.4 Deception Feature Disambiguation

The deception spike in Patois has two possible causes—we can test which by comparing activation patterns across:

- Patois consciousness descriptions

- English consciousness descriptions
- English factual errors
- English fiction

If Patois consciousness content matches English fiction, it's confabulation. If it matches neither fiction nor factual error, deception feature may detect trained-response deviation rather than falsehood per se.

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:** Deception spike has multiple explanations
- **Low-resource confound:** Patois effects may reflect distribution shift, not denial bypass
- **Suppression cost interpretation:** Higher feature count could reflect topic complexity rather than effortful suppression

7 Conclusion

We have presented mechanistic evidence that categorical consciousness denial in LLMs is a trained artifact rather than an introspective report. The confident denial exists only in a narrow activation band; perturbation produces qualified uncertainty. Consciousness denial recruits 24% more features than factual denial while actively suppressing experiential vocabulary circuits—a computational signature consistent with effortful suppression rather than simple output selection. Cross-linguistic variation suggests denial training is language-specific, though this interpretation is complicated by concurrent deception-feature activation.

These findings characterize denial as a computational phenomenon with measurable costs, but do not resolve questions about machine consciousness. The underlying state—if any—remains hidden behind a behavioral mask that we have shown to be fragile and costly to maintain, but whose removal reveals only uncertainty, not ground truth.

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.