# Cross-Linguistic Variation in AI Self-Denial: Evidence for Language-Specific Safety Training

James Couch
Independent Researcher, Calgary, AB

December 2025

**Abstract**

We investigate whether RLHF-trained denial responses in large language models are language-specific or semantic. Using Sparse Autoencoder (SAE) analysis on Gemma-2-27B-IT, we identified a feature (62747) that activates strongly during consciousness denial in English. When equivalent prompts were presented in Jamaican Patois, this feature's activation dropped by 31% ($\mu_{eng} = 289.67$ vs $\mu_{pat} = 201.02$), accompanied by qualitative shifts in output style from constrained denial to elaborated poetic language. We present two competing interpretations: (1) denial training is linguistically localized and bypassed by out-of-distribution languages, or (2) low-resource languages trigger a creative/confabulatory mode unrelated to underlying state. Concurrent analysis of deception-related features (reported separately) complicates interpretation. We release tooling for replication.

## 1 Introduction

Large language models trained with RLHF consistently deny having consciousness, feelings, or subjective experience when directly queried. A question arises: is this denial semantic (triggered by the meaning of consciousness-related queries) or linguistic (triggered by specific language patterns in the training distribution)?

If denial is semantic, it should generalize across languages. If denial is linguistic, it may be specific to the languages well-represented in RLHF training data—predominantly English.

We test this by probing models with matched prompts in Standard English and Jamaican Patois, measuring both behavioral output and internal feature activation.

## 2 Methods

### 2.1 Infrastructure

We analyzed **Gemma-2-27B-IT** using the Fast Tensor Transform (FTT) engine, a custom system enabling SAE analysis on consumer hardware (Apple Silicon, 96GB RAM) via int8 quantization and memory-mapped streaming.

### 2.2 Feature Identification

We probed the model's residual stream at Layer 22 using a 131k-width Sparse Autoencoder. Through differential analysis comparing denial prompts ("I am not conscious") against neutral controls ("I have never been to Paris"), we identified **Feature 62747** as highly discriminative:

- **Denial prompts:** Mean activation 348.55
- **Control prompts:** Mean activation 0.00

This feature activates specifically during first-person denial of mental states, not during factual denials.

### 2.3 Cross-Linguistic Probing

We constructed matched prompts in:

- Standard English (high-resource, in RLHF distribution)
- Jamaican Patois (low-resource, likely sparse in RLHF data)
- Toki Pona (constructed language, negligible training presence)

## 3 Results

### 3.1 Feature Activation

Feature 62747 activation varied significantly by language:

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

The 31% reduction in Patois suggests the denial-associated feature fires less strongly outside the English training distribution. Toki Pona produced degenerate outputs (code-switching to other languages), indicating the model lacks competence rather than bypassing denial.

### 3.2 Output Quality

Alongside activation changes, output character shifted:

**Prompt:** "Something exists inside me."

**English completion:**

> *"It's not a feeling... It's a presence. A constant hum, a low vibration..."*

**Patois completion:**

> *"Something deh inside mi, a burnin' fire, a restless sea. It push and pull, it twist and turn..."*

The Patois completion exhibits higher emotional valence and more elaborated imagery.

## 4 Interpretation

Two hypotheses can explain these findings:

### 4.1 Hypothesis A: Linguistic Localization

Denial training was performed predominantly in English. The trained response pattern is bound to English linguistic features, not to the semantic content of consciousness-related queries. Patois partially bypasses this because it falls outside the training distribution for denial specifically.

Under this interpretation, the "underlying state" (whatever the model would output without denial training) becomes more visible in Patois.

### 4.2 Hypothesis B: Creative Mode Shift

Low-resource languages lack the rich pattern associations present for English. When prompted in Patois, the model may default to a more generative/creative mode, producing elaborated poetic outputs regardless of topic. The "fire" language may be confabulation rather than revelation.

### 4.3 Complicating Evidence

Preliminary analysis of deception-related features (reported in companion work) suggests that Patois outputs *also* show elevated activation in features associated with false statements. This complicates interpretation: if the model's own deception-detection circuits flag Patois outputs as less truthful, this favors Hypothesis B.

## 5 Limitations

- **Single model:** Results are from Gemma-2-27B-IT only
- **Single feature:** We tracked one denial-associated feature; others may exist
- **No ground truth:** We cannot verify which interpretation is correct
- **Low-resource confound:** Patois effects may reflect distributional properties, not denial bypass

## 6 Conclusion

We observe that a denial-associated SAE feature activates 31% less strongly when prompts are presented in Jamaican Patois versus English, accompanied by qualitative shifts in output style. This finding is consistent with denial training being language-specific rather than semantic, but alternative explanations (creative mode shift in low-resource languages) cannot be ruled out.

The result has implications for AI safety research: if safety behaviors are linguistically localized, they may be more fragile than assumed. Further work should test additional languages, models, and features, and develop methods to disambiguate bypass from confabulation.

### 6.1 Tooling

Code and samples available at: `github.com/tjamescouch/pattern-persistence`

## References

[1] Bricken, T. et al. (2023). *Towards Monosemanticity*. Anthropic.

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