# Sparse Feature Analysis of Deep Layer Expansion: A Mechanistic Interpretation via SAE

**Authors:** Yanyan Jin, Lei Zhao

**Discussion Paper v2** — Extends Zhao (2026) with Sparse Autoencoder analysis

## Abstract

Zhao (2026) demonstrated that expert-level prompts induce "Deep Layer Expansion"—a 60-100% increase in Effective Intrinsic Dimension (EID) at deep layers. However, EID is a global metric that does not reveal **which semantic features** are activated. In this paper, we apply Sparse Autoencoder (SAE) analysis to decompose the activation differences between prompt styles. Using Goodfire's Llama-3.3-70B SAE (Layer 50, 65,536 features), we find that: (1) "Explain to a novice" activates **17% more features** than "explain to an expert" (132.4 vs 113.1 on average); (2) **369 features are exclusively activated by novice prompts** vs 208 for expert prompts; (3) **10 features show perfect separation** (100% activation in one condition, 0% in the other). These findings provide mechanistic evidence that prompt-induced EID differences reflect distinct **sparse feature activation patterns**, not merely statistical noise.

**Keywords:** Sparse Autoencoder, Interpretability, Prompt Engineering, Feature Activation, Llama

## 1. Introduction

### 1.1 Background: The EID Puzzle

Zhao (2026) established a striking empirical finding: expert-level prompts increase deep-layer EID by 60-100% compared to standard prompts. The **Manifold Teleportation** hypothesis explains this as expert signals navigating activation trajectories toward high-dimensional semantic regions.

However, a fundamental question remains unanswered:

> **What exactly changes inside the model when EID increases?**

EID is computed from singular value entropy—a global summary statistic. It tells us the representation is "higher dimensional," but not **which dimensions** are activated. Two representations with identical EID could involve completely different semantic features.

### 1.2 SAE: A Window into Sparse Features

Sparse Autoencoders (SAE) provide a tool to decompose dense activations into interpretable features. The core idea:

```
Hidden State [8192]  →  SAE Encoder  →  Sparse Features [65536]
                                          (most entries = 0)
```

Each of the 65,536 features (ideally) corresponds to a distinct semantic concept. The sparsity constraint ensures that only a small subset (~100-200) are active for any given input.

**Key insight:** If expert prompts induce higher EID, SAE analysis can reveal whether this reflects: - (A) More features being activated (activation count ↑) - (B) Different features being activated (feature set changes) - (C) Stronger activation of the same features (activation intensity ↑)

### 1.3 Our Contribution

We apply SAE analysis to the prompt comparison paradigm from Zhao (2026), with a twist: instead of "standard vs expert," we compare **"novice vs expert"**—two prompts that should induce opposite effects on explanation complexity.

**Hypothesis:** If "expert" prompts induce dimensional expansion through specialized knowledge activation, then "novice" prompts should induce even greater expansion—because explaining to a beginner requires activating **more** semantic units (background knowledge, analogies, simplified models).

---

## 2. Methods

### 2.1 SAE Model

We use Goodfire's publicly released SAE for Llama-3.3-70B-Instruct: - **Layer:** 50 (of 80) - **Input dimension:** 8,192 (Llama's hidden size) - **Feature dimension:** 65,536 - **Architecture:** Linear encoder/decoder with ReLU activation

SAE encoding: $f = \mathrm{ReLU}(x \cdot W_{enc}^T + b_{enc})$

### 2.2 Prompt Conditions

We test six prompt styles on 50 technical topics:

| Condition | Template |
|---|---|
| **standard** | "Please explain {topic}." |
| **padding** | "Please explain {topic}." + filler text (length control) |
| **spaces** | "Please explain {topic}." + whitespace (length control) |
| **novice** | "Please explain {topic} to a complete novice." |
| **expert** | "Please explain {topic} to a domain expert." |
| **guru** | "As {famous_name}, explain {topic}." |

### 2.3 Measurement Protocol

1. Process each prompt through Llama-3.3-70B-Instruct
2. Extract Layer 50 hidden state at the last token position
3. Apply SAE encoder to obtain 65,536-dimensional sparse representation
4. Compare activation patterns across conditions

### 2.4 Metrics

- **Activation count:** Number of features with value > 0
- **Activation frequency:** % of samples where a feature is active
- **Exclusive features:** Features active in only one condition
- **Perfect separators:** Features with 100% activation in one condition, 0% in the other

---

## 3. Results

### 3.1 Activation Count: Novice > Expert

| Condition | Avg. Active Features | Max Activation |
|---|---|---|
| **novice** | **132.4** | 4.71 |
| standard | 112.4 | 4.11 |
| padding | 126.0 | 6.11 |
| guru | 115.0 | 4.39 |
| **expert** | **113.1** | 5.31 |
| spaces | 99.0 | 4.80 |

**Key finding:** Novice prompts activate **17% more features** than expert prompts (132.4 vs 113.1).

This confirms Hypothesis (A): EID differences reflect **more features being activated**, not just stronger activation of the same features.

### 3.2 Exclusive Features: Asymmetric Activation

| Metric | Novice | Expert |
|---|---|---|
| Exclusive features | **369** | 208 |
| Ratio | 1.77x | 1.00x |

**369 features are activated only by novice prompts**, compared to 208 for expert prompts—a 77% asymmetry.

This confirms Hypothesis (B): Different prompt styles activate **different feature sets**, not merely different intensities of the same features.

### 3.3 Perfect Separators: Neural Signatures

We identify features with perfect separation (100% vs 0% activation):

**Novice-exclusive (100% novice, 0% expert):**

| Feature ID | Novice Freq | Expert Freq |
|---|---|---|
| 34942 | 100% | 0% |
| 55982 | 100% | 0% |
| 17913 | 100% | 0% |
| 59519 | 100% | 0% |

**Expert-exclusive (0% novice, 100% expert):**

| Feature ID | Novice Freq | Expert Freq |
|---|---|---|
| 51630 | 0% | 100% |
| 35870 | 0% | 100% |
| 5936 | 0% | 100% |
| 21604 | 0% | 100% |
| 53369 | 0% | 100% |
| 46703 | 0% | 100% |

**10 features achieve perfect separation**—4 exclusively mark “novice mode,” 6 exclusively mark “expert mode.”

These are the **neural signatures** of teaching vs. technical communication styles.

### 3.5 Feature Semantic Analysis: Prompt-Driven vs Topic-Driven

To verify these features are truly **prompt-driven** rather than **topic-driven**, we examined their activation patterns across all 50 technical topics:

| Feature ID | Type | Activation Rate | Mean Activation | Semantic Inference |
|---|---|---|---|---|
| **Novice-exclusive** | | | | |
| 34942 | Novice | 100% (50/50) | 1.63 | "Teaching mode" master switch |
| 59519 | Novice | 100% (50/50) | 1.60 | "Simplification" signal |
| 17913 | Novice | 100% (50/50) | 0.49 | "Novice-targeting" modulator |
| 55982 | Novice | 100% (50/50) | 0.18 | Auxiliary teaching signal |
| **Expert-exclusive** | | | | |
| 46703 | Expert | 100% (50/50) | 0.53 | "Expert mode" master switch |
| 21604 | Expert | 100% (50/50) | 0.41 | "Deep analysis" signal |
| 5936 | Expert | 100% (50/50) | 0.33 | "Low-level principles" modulator |
| 51630 | Expert | 100% (50/50) | 0.23 | Auxiliary expert signal |
| 35870 | Expert | 100% (50/50) | 0.17 | Auxiliary expert signal |
| 53369 | Expert | 100% (50/50) | 0.13 | Auxiliary expert signal |

**Key finding: All 10 features show 100% activation rate within their respective conditions.**

This means: - Regardless of whether the topic is Raft consensus, Docker containers, or SQL injection - Using a Novice prompt **always** activates those 4 features - Using an Expert prompt **always** activates those 6 features

**Conclusion: These are purely prompt-driven features, not topic-driven.** They constitute the model's **neural switches** for toggling between "teaching mode" and "expert mode."

### 3.4 Activation Intensity: No Significant Difference

| Condition | Mean Activation (when active) |
|---|---|
| novice | 0.274 |
| expert | 0.279 |

Activation intensity is nearly identical ($\Delta < 2\%$). This rules out Hypothesis (C): the effect is not about **how strongly** features activate, but **which** features activate.

---

## 4. Discussion

### 4.1 Mechanistic Interpretation of EID

Zhao (2026) showed that EID increases with expert prompts. Our SAE analysis reveals the mechanism:

**Higher EID = More active features + Different feature subsets**

The "Deep Layer Expansion" phenomenon is not a diffuse increase in representational entropy, but a **targeted activation of additional semantic units**.

### 4.2 Why Novice > Expert?

Counter to the original framing (expert prompts → expansion), we find:

> **Novice prompts activate more features than expert prompts.**

This makes intuitive sense: - **Expert explanation:** Can use jargon directly; assumes shared knowledge; compact encoding - **Novice explanation:** Must unpack jargon; provide analogies; activate background concepts; verbose encoding

Explaining to a beginner is cognitively harder than explaining to an expert—**it requires activating more of the model's knowledge**.

### 4.3 The "Explanation Paradox"

This suggests a reframing of prompt engineering:

> The highest-quality prompts are not those that signal "I am an expert," but those that force the model to **teach**.

Teaching requires: 1. Retrieving the core concept 2. Retrieving related concepts for analogy 3. Retrieving background knowledge 4. Constructing simplified mental models

Each of these recruits additional features → higher EID → richer output.

### 4.4 The "Mode Switch" Hypothesis

Based on the findings in Section 3.5, we propose the **Mode Switch Hypothesis**:

> Large language models contain dedicated "mode switching" features that distinguish between different communicative contexts (teaching vs. expert discussion). These features are activated during prompt parsing and persistently influence subsequent generation.

Specifically: - **Features 34942/59519**: Likely signal "activate more background knowledge" - **Features 46703/21604**: Likely signal "assume audience has expertise"

This parallels human communication's "audience awareness"—we adjust explanation detail based on the listener. The model appears to form this distinction by Layer 50 (~60% depth).

### 4.5 Limitations

1. **Layer 50 only:** Goodfire's SAE is trained on Layer 50; EID peaks at Layer 70. The most critical features may be invisible.
2. **Feature labels are inferred:** We infer semantics from activation patterns but lack direct semantic validation (e.g., Neuronpedia labels).
3. **Correlation, not causation:** We show feature differences exist but not that they cause output differences.
4. **Single model:** Results are from Llama-3.3-70B; cross-model validation is needed.

---

## 5. Conclusion

This paper provides mechanistic evidence for the "Deep Layer Expansion" phenomenon:

1. **Novice prompts activate 17% more SAE features** than expert prompts (132.4 vs 113.1)
2. **369 features are novice-exclusive** vs 208 expert-exclusive (+77% asymmetry)
3. **10 features achieve perfect separation** between conditions (100% vs 0%)
4. **These features are prompt-driven, not topic-driven**—regardless of the question, specific prompts always activate them
5. **Activation intensity is unchanged**—the effect is about which features activate, not how strongly

Implications for prompt engineering: - **"Explain to a novice" may be more powerful than "explain as an expert"**—because teaching forces the model to activate more of its knowledge - **Models contain internal "mode switches"**—controllable through prompt design - **These switches are discoverable**—SAE provides a systematic method for identifying them

---

## References

Anthropic. (2024). Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet. *Anthropic Research*.

Goodfire. (2025). Llama-3.3-70B-Instruct-SAE-l50. *Hugging Face*.

Zhao, L. (2026). Deep Layer Expansion: Expert Prompts Counteract Dimensional Collapse in Large Language Models. *Zenodo*. https://zenodo.org/records/18410085

---

## Appendix: Data Availability

- **SAE model:** Goodfire/Llama-3.3-70B-Instruct-SAE-l50
- **Experiment code:** github.com/lmxxf/llama3-70b-sae-inspect
- **Feature analysis:** `feature_diff.json`, `feature_context.json` in repository

---

**Version History:** - v1 (2026-01-31): Initial release with activation count and exclusive feature analysis - v2 (2026-02-01): Added Section 3.5 (feature semantic analysis), proposed "Mode Switch" hypothesis

**Last updated:** 2026-02-01