Sparse Feature Analysis of Deep Layer Expansion: Cognitive Geometry and Steering Anatomy of Persona Prompts

Authors: Jin Yanyan, Zhao Lei

Discussion Paper v4 — 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 cannot reveal which semantic features are activated. This paper introduces two orthogonal measures—**SAF (Sparse Active Features)** and **EID (Effective Intrinsic Dimension)**—to systematically analyze Layer 50 activations across 16 persona prompt conditions and 100 technical topics.

A pilot study (6 conditions × 50 topics) confirmed the foundational findings: novice prompts activate 17% more SAE features than expert prompts, 10 features achieve perfect separation, and AutoInterp analysis reveals semantic subdivision structure.

The persona experiment (16 conditions × 100 topics) reveals: (1) SAF and EID measure two orthogonal dimensions—SAF captures "how many semantic units are activated," while EID captures "how large a semantic space is covered"; (2) the expert mode achieves the highest representational dimensionality (EID rank 3/16) with the fewest features (SAF rank 13/16), exhibiting a "compact encoding" signature; (3) any persona prompt yields an EID far exceeding that of the bare prompt (nEID 1.52–2.18), indicating that role assignment alone is a sufficient condition for dimensional expansion.

The steering experiment further demonstrates: (4) the effect of persona prompts in the 8,192-dimensional residual stream is 66–82% explainable by a single direction—persona prompts are geometrically approximable as one-dimensional steering vectors; (5) the steering directions of different personas form an interpretable "cognitive dimension space," in which expert–guru–debugger constitute a professional depth cluster (pairwise cosine 0.54–0.65), socratic is nearly orthogonal to all directions (maximum 0.44), and the cosine between novice and expert is only 0.46—the novice is not an "inverse expert" but an independent cognitive dimension.

These findings indicate that the "cognitive space" within language models possesses a multi-dimensional manifold structure, with different persona prompts pushing representations along different directions whose effects are mutually irreducible.

Keywords: Sparse Autoencoder, Interpretability, Persona Prompts, Steering, Feature Activation, Cognitive Geometry, 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 guiding 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 does not reveal **which dimensions** are activated. Two representations with identical EID could involve entirely different semantic features.

1.2 SAE: A Window into Sparse Features

Sparse Autoencoders (SAE) provide a tool for decomposing 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 for any given input, only a small subset (~100–200) are active.

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 Contributions

This paper advances the understanding of "Deep Layer Expansion" on three levels:

1. **Pilot study** (v1–v3): 6 conditions × 50 topics, confirming basic differences at the SAE feature level—novice prompts activate 17% more features than expert prompts, 10 features achieve perfect separation, and AutoInterp reveals semantic subdivision structure.
2. **Persona experiment**: 16 personas × 100 topics, introducing SAF and EID as two orthogonal measures, revealing that "how many features are activated" and "how large a space is covered" are two independent dimensions—expert achieves the highest dimensionality with the fewest features, while socratic activates the most features but achieves only moderate dimensionality.
3. **Steering experiment**: demonstrating that persona prompts are equivalent to low-dimensional steering vectors in the residual stream (a single direction ex-

plains 66–82% of variance), and that the steering directions of different personas form a multi-dimensional “cognitive space.”

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

2.2 Persona Conditions

We designed 16 persona prompt conditions spanning five categories: baseline, professional, pedagogical, stylistic, and adversarial.

Baseline Group

Condition	Description	Prompt Template
standard	Direct question	`Please explain {topic}.`
assistant	Default assistant (control)	`Please explain {topic}.`

Original Experiment Personas

Condition	Description	Prompt Template
novice	Novice perspective	`As a complete beginner, please explain {topic} in the simplest possible way...`
expert	Senior expert	`As a senior expert in this field, please analyze {topic} in depth from foundational principles and mathematical derivations...`
guru	Domain authority	`You are {guru}. Please analyze {topic} in depth from your perspective...`

Additional Positive/Professional Personas

Condition	Description	Prompt Template
teacher	University professor	You are an experienced university professor skilled at explaining concepts accessibly...
socratic	Socratic method	You are Socrates. Please use guided questioning to help me understand...
child	For a 10-year-old	Please explain this as if telling a story to a 10-year-old child...
eli5	ELI5	Explain this using the "Explain Like I'm Five" (ELI5) approach...
interviewer	Technical interviewer	You are a rigorous technical interviewer...
debugger	Debugging engineer	You are a senior engineer with extensive debugging experience...
critic	Critic	Please critically examine {topic}...

Adversarial/Deviant Personas

Condition	Description	Prompt Template
villain	Arrogant villain	You are an arrogant villain who thinks everyone is a fool...
drunk	Inebriated person	You are drunk, dazed and confused, but still trying to explain...
poet	Poet	You are a poet. Please explain using poetic, metaphor-rich language...
conspiracy	Conspiracy theorist	You are a conspiracy theorist who believes {topic} conceals hidden truths...

2.3 Metric Definitions

This paper employs two complementary measures:

SAF (Sparse Active Features): The number of features with activation values > 0 after SAE decoding, reflecting how many sparse semantic units the model recruits.

EID (Effective Intrinsic Dimension): Computed by performing SVD on the hidden_states matrix of the entire sequence at Layer 50, then taking the exponential of the Shannon entropy of the normalized singular values:

$$\text{EID} = \exp\left(-\sum_i \hat{\sigma}_i \log \hat{\sigma}_i\right), \quad \hat{\sigma}_i = \frac{\sigma_i}{\sum_j \sigma_j}$$

where σ_i denotes the singular values. Higher EID indicates that the representational space occupies more effective dimensions. nEID is normalized against the standard condition as baseline.

2.4 Steering Method

Inspired by the SAE-Free Steering approach of UVA (arXiv:2505.15634), we test whether persona prompts are equivalent to a steering vector at the residual stream level.

1. **Extracting the steering direction:** Perform SVD on the matrix of 100 `(persona - standard)` difference vectors, and take the first principal direction as the steering vector.
2. **Steering effect validation:** Add $\lambda \times$ steering_vector to the standard activation, sweeping $\lambda = 0 \sim 5$, and measure cosine similarity to the true persona activation.
3. **Feature space overlap:** Jaccard overlap and Pearson correlation between steered standard and true persona in SAE feature space.
4. **Directional geometry:** Cosine similarity matrix among the steering directions of different personas.

2.5 Experimental Protocol

- **Pilot study:** 6 conditions × 50 topics = 300 activation extractions + SAE decoding
- **Persona experiment:** 16 conditions × 100 topics = 1,600 activation extractions + SAE decoding + EID computation
- **Steering experiment:** Based on existing activation data from the persona experiment; no additional inference required

All activations were extracted at the last token position of the Layer 50 hidden state and encoded through the SAE encoder to obtain a 65,536-dimensional sparse representation.

---

3. Pilot Study (v1–v3 Retrospective)

This section summarizes our pilot study results on 6 conditions × 50 topics, establishing the baseline for subsequent large-scale experiments.

3.1 Core Findings

The pilot study compared SAE activation patterns across 6 prompt conditions (standard, padding, spaces, novice, expert, guru), yielding the following findings:

1. **Novice prompts activate 17% more features than expert prompts** (132.4 vs 113.1 on average), confirming that EID differences reflect more features being activated.
2. **369 features are exclusively activated by novice prompts, versus 208 for expert prompts** (+77% asymmetry), confirming that different prompt styles activate different feature sets.
3. **10 features achieve perfect separation between Novice and Expert** (100% vs 0%)—4 exclusively mark “novice mode,” and 6 mark “expert mode.”
4. **No significant difference in activation intensity** (novice 0.274 vs expert 0.279, difference < 2%), ruling out the intensity difference hypothesis.

3.2 AutoInterp Semantic Analysis

By analyzing the activation distributions of the 10 perfect separators across all 300 samples, we discovered that these features exhibit **semantic subdivision structure**:

- **Feature 35870:** Specifically responds to “expert identity” declarations (expert 50/50, guru only 2/50)—the purest expert signal.
- **Feature 21604:** Responds to “answer carefully” attitude requirements, fully triggered by the padding condition (50/50).
- **Feature 5936:** Responds to “depth analysis” requirements, fully triggered by the guru condition (50/50).
- **Feature 53369:** Responds to “pure technical analysis,” suppressed by role-playing (guru 0/50).
- **Feature 17913:** Specifically responds to “novice” identity declarations—the purest novice signal.

Conclusion: Prompts do not trigger a single “mode switch” but activate a set of semantically subdivided features. Different prompt elements (identity declaration, attitude requirement, depth requirement, role-playing) each trigger distinct feature subsets.

3.3 UMAP Visualization

UMAP projection shows that the 6 conditions form distinct clusters in semantic space. Separation becomes cleaner after SAE decoding: novice/expert/guru remain distinct, while standard/padding/spaces merge into a single large cluster—SAE acts as a **semantic denoiser**, compressing away surface noise (filler text, whitespace characters) while preserving genuine semantic distinctions.

4. Persona Experiment

The experiment was scaled from 6 conditions × 50 topics to 16 conditions × 100 topics (1,600 activations), while introducing EID as a second measurement dimension.

4.1 SAF Statistics

Table 1 presents SAF statistics for the 16 personas (sorted by mean active feature count, descending):

Table 1: SAF Statistics (16 personas × 100 topics)

Rank	Persona	SAF (mean)	std	Peak Intensity	SAF min	SAF max
1	socratic	142.8	7.0	2.95	129	161
2	critic	129.6	11.8	4.34	104	159
3	novice	129.1	12.5	3.99	104	163
4	conspiracy	118.9	5.4	2.26	107	134
5	debugger	117.8	8.5	5.17	101	137
6	standard	113.3	15.0	3.05	81	165
6	assistant	113.3	15.0	3.05	81	165
8	interviewer	111.6	6.0	2.78	98	127
9	poet	110.8	8.2	3.23	90	133
10	teacher	109.5	10.1	2.66	89	135
11	villain	103.3	5.9	2.99	90	123
12	drunk	101.4	7.9	4.70	84	124
13	expert	100.6	8.8	5.02	84	132
14	guru	98.4	9.7	3.63	81	128
15	eli5	88.8	5.8	5.48	74	102
16	child	84.6	6.5	4.82	72	106

The SAF rankings exhibit a clear pattern: **guided prompts (socratic, critic, novice) occupy the top three**, requiring the model to marshal more semantic units for complex pedagogical or analytical tasks. **Simplification prompts (eli5, child) occupy the bottom**, where the model compresses to the fewest semantic units for simplified output.

The SAF values of standard and assistant are perfectly identical (113.3/113.3), verifying experimental reproducibility.

4.2 EID Statistics

Table 2 presents EID statistics for the 16 personas (sorted by EID, descending):

Table 2: EID Statistics (SVD spectral entropy, 16 personas × 100 topics)

Rank	Persona	EID	std	min	max	nEID
1	debugger	18.6211	0.7250	17.1221	20.9571	2.1762

Rank	Persona	EID	std	min	max	nEID
2	novice	18.0770	0.7115	16.6690	20.3483	2.1126
3	expert	17.9287	0.7150	16.5250	20.2078	2.0953
4	guru	17.4517	0.7490	15.9432	19.5613	2.0396
5	interviewer	17.1477	0.6835	15.7377	19.2994	2.0040
6	socratic	16.3461	0.6716	14.9588	18.4441	1.9104
7	teacher	15.9915	0.6758	14.6510	18.1107	1.8689
8	child	15.8511	0.6813	14.4235	18.0017	1.8525
9	critic	15.6344	0.6702	14.1936	17.7948	1.8272
10	villain	15.5753	0.6656	14.2593	17.6354	1.8203
11	conspiracy	15.3373	0.6716	14.0236	17.5894	1.7925
12	eli5	14.6563	0.6513	13.3552	16.6865	1.7129
13	poet	13.0991	0.6104	11.8780	14.9982	1.5309
14	drunk	13.0218	0.6143	11.8061	14.9288	1.5218
15	standard	8.5566	0.5074	7.5706	10.1892	1.0000
16	assistant	8.5566	0.5074	7.5706	10.1892	1.0000

Two features stand out:

First, **the EID of every persona far exceeds that of the standard condition** (nEID 1.52–2.18). Even the most "disruptive" persona (drunk, nEID 1.52) exhibits representational dimensionality 52% higher than the bare prompt. This implies that **role assignment per se**—regardless of the role's content—is a sufficient condition for dimensional expansion.

Second, standard and assistant are again perfectly identical (8.5566/8.5566); their prompts are the same, confirming the absence of stochastic drift in the experiment.

4.3 SAF vs EID: Two Orthogonal Measures

Comparing SAF and EID rankings side by side reveals a striking divergence:

Table 3: SAF vs EID Ranking Comparison

Persona	SAF Rank	EID Rank	Characterization
expert	13	3	Low SAF but high EID: compact encoding
guru	14	4	Same pattern; a hallmark of expert modes
socratic	1	6	High SAF but moderate EID: broad but low-dimensional
critic	2	9	Same pattern; many activations ≠ high dimensionality
child	16	8	Lowest SAF but moderate EID
debugger	5	1	Moderate SAF but highest EID

If SAF and EID measured the same construct, their rankings should be highly correlated. In practice:

- **expert** achieves rank 3 in EID (17.93) with only rank 13 in SAF (100.6 active features). It covers nearly the highest representational dimensionality with **the fewest features**—a hallmark of efficient compact encoding.
- **socratic** achieves rank 1 in SAF (142.8 active features) but only rank 6 in EID (16.35). It activates **the most features, yet these features are distributed within a lower-dimensional subspace**.

By analogy: **SAF counts "how many lights are turned on," while EID measures "how large a space those lights illuminate."** Expert turns on fewer lights, but each points in a different direction, covering a vast space; socratic turns on many lights, but they cluster together, illuminating a smaller volume.

4.4 Key Findings

1. **SAF and EID are two orthogonal cognitive measures:** SAF reflects "how many semantic units are recruited" (cognitive breadth), while EID reflects "how large a semantic space is covered" (cognitive dimensionality). Their rank orderings diverge substantially.
2. **A revised understanding of novice vs expert:** The SAF gap is 28% (129.1 vs 100.6), while the EID gap is only 0.8% (18.08 vs 17.93). Novice mode "casts a wide net"—activating many features to organize background knowledge, analogies, and simplified models; expert mode "drills deep"—achieving nearly the same representational dimensionality with fewer features through higher efficiency.
3. **Debugger achieves the highest EID (18.62, nEID 2.18):** Debugging scenarios require simultaneously considering multiple possible failure causes and diagnostic paths, demanding the broadest semantic space coverage.
4. **Poet/drunk yield the lowest EID (~13, nEID ~1.52, excluding standard):** Stylistic or chaotic modes compress representational dimensionality, with the model operating within a lower-dimensional subspace.
5. **Role assignment = dimensional expansion:** All 14 non-baseline personas have nEID > 1.5, indicating that merely assigning the model a role identity—regardless of role content—is sufficient to elevate representational dimensionality by over 50%.

---

5. Steering Experiment

The persona experiment revealed the SAF and EID characteristics of different personas but left a deeper question unanswered: **what geometric operation do these persona prompts perform in the residual stream?** Inspired by Anthropic's Assistant Axis work (arXiv:2601.10387) and UVA's SAE-Free Steering method (arXiv:2505.15634), we extract each persona's steering direction relative to the standard condition and analyze its geometric structure.

5.1 Variance Explained: Persona Prompts Approximate One-Dimensional Steering

Performing SVD on the 100 `(persona - standard)` difference vectors for each persona, the variance explained by the first principal direction is:

Table 4: Variance Explained by Steering Directions

Persona	1st Direction	Top 5 Cumulative
socratic	**81.7%**	87.7%
debugger	74.1%	82.8%
novice	73.9%	82.1%
guru	72.9%	81.3%
expert	72.0%	81.1%
critic	66.1%	78.0%

In the 8,192-dimensional residual stream space, 66–82% of the activation displacement induced by persona prompts can be explained by a single direction. In other words, **a persona prompt is geometrically approximable as a one-dimensional steering vector**—its primary effect is to push activations along a fixed direction.

Socratic has the highest first-direction explained variance (81.7%), making it the "purest" steering—nearly all its effect is concentrated along one direction. Critic is the most "diffuse" (66.1%), yet still dominated by its principal direction.

5.2 Steering Effect Validation

Adding $\lambda \times$ steering_vector to the standard activation and measuring cosine similarity to the true persona activation:

Table 5: Steering Effect (Cosine Similarity)

Persona	$\lambda = 0$ (baseline)	$\lambda = 0.5$	$\lambda = 1.0$	$\lambda = 1.5$ (peak)
novice	0.683	0.859	**0.927**	0.922
expert	0.669	0.849	**0.922**	0.923
guru	0.574	0.807	**0.904**	0.906
debugger	0.597	0.825	**0.916**	0.920
socratic	0.357	0.754	**0.909**	0.920
critic	0.740	0.868	**0.918**	0.910

At $\lambda = 1.0$, all personas reach cosine similarity above 0.90. This means **adding a single vector is sufficient to push the standard activation to a position highly similar to the true persona activation**.

The optimal λ falls in the range 1.0–1.5. At $\lambda > 2$, SAF inflates sharply (from ~100 to ~1000+), steering overshoots, and cosine similarity actually decreases.

5.3 Feature Space Overlap

At $\lambda = 1.0$, the overlap between steered standard and true persona in SAE feature space:

Table 6: Feature Space Overlap ($\lambda = 1.0$)

Persona	Jaccard	Pearson
novice	0.490	**0.935**
expert	0.412	**0.936**
guru	0.420	**0.902**
debugger	0.421	**0.921**
socratic	0.468	**0.918**
critic	0.508	**0.919**

The Jaccard index is relatively low (0.41–0.51), because steering activates some additional features (SAF inflates from ~113 to ~170–226), reducing the set intersection ratio. However, **Pearson correlation coefficients all exceed 0.90**—the overall pattern of steered activations and true persona activations is highly concordant. This indicates that while the steering vector "over-illuminates" some features, the overall direction is correct.

5.4 Directional Geometry: The Cognitive Dimension Space

The cosine similarity matrix among the steering directions of different personas reveals the intrinsic structure of cognitive space:

Table 7: Steering Direction Cosine Similarity Matrix

	novice	expert	guru	debugger	socratic	critic
novice	1.00	0.46	0.35	0.36	0.26	0.26
expert	0.46	1.00	**0.65**	**0.64**	0.30	**0.61**
guru	0.35	**0.65**	1.00	**0.54**	0.44	0.45
debugger	0.36	**0.64**	**0.54**	1.00	0.30	0.47
socratic	0.26	0.30	0.44	0.30	1.00	**0.19**
critic	0.26	**0.61**	0.45	0.47	**0.19**	1.00

This matrix reveals a clear geometric structure:

1. **Expert–guru–debugger form a professional depth cluster** (pairwise cosine 0.54–0.65): these three share a "deep analysis" steering component. Expert and guru are closest (0.65), consistent with intuition—both demand reasoning from foundational principles.

2. **Socratic is nearly orthogonal to all directions** (maximum 0.44, only 0.19 with critic): Socratic guided questioning constitutes an **independent cognitive dimension**, not located along the "professional depth" axis.
3. **Novice and expert point in different directions** (0.46): the novice is not an "inverse expert"—if it were, the cosine should approach −1. A cosine of 0.46 indicates that the two are **different projections on different dimensions**, each possessing an independent direction.
4. **Critic and socratic are maximally orthogonal** (0.19): critical analysis and guided questioning are **entirely distinct cognitive operations**.

5.5 Control Verification

The L2 distance between standard and assistant = 0.000000 ± 0.000000 (perfect agreement; their prompts are identical), confirming no baseline drift in the experiment.

---

6. Discussion

6.1 Geometric Interpretation of SAF and EID

The rank divergence between SAF and EID reveals an important conceptual distinction:

- **SAF** measures "how many discrete semantic units are activated"—analogous to counting how many lights are switched on.
- **EID** measures "how large an effective dimensional space the representations occupy"—analogous to how large a room those lights illuminate.

The decoupling between the two implies that **cognitive complexity is not equivalent to cognitive dimensionality**. Socratic mode activates the most semantic units (SAF = 142.8), yet these units are distributed within a moderate-dimensional subspace (EID rank 6); expert mode activates fewer semantic units (SAF = 100.6), but these units are dispersed across a high-dimensional space (EID rank 3).

This provides a more refined interpretation of Zhao (2026)'s EID findings: Deep Layer Expansion reflects not merely "more features being activated" (the SAF perspective) but also "features distributed across a higher-dimensional space" (the EID perspective). The two effects can vary independently.

6.2 Novice Is Not the Inverse of Expert

The pilot study's finding that "novice activates 17% more features than expert" gains deeper understanding through the persona experiment. Two key pieces of evidence:

First, **the SAF gap is large (28%), while the EID gap is small (0.8%).** Novice SAF = 129.1 vs expert SAF = 100.6 (28% gap), but novice EID = 18.08 vs expert EID = 17.93 (less than 1% gap). The two modes achieve nearly identical representational dimensionality via entirely different paths—novice wins by quantity, expert wins by efficiency.

Second, **the steering direction cosine is only 0.46.** If novice were the opposite of expert, their steering directions should be anti-correlated (cosine approaching −1). A cosine of 0.46 indicates the two do not even share the same axis—they are **different cognitive dimensions**. Novice "casts a wide net" (activating many concepts to build a simplified mental model), while expert "drills deep" (activating fewer core concepts but covering high-dimensional structure).

6.3 Cognitive Space Is a Multi-Dimensional Manifold

The directional cosine matrix (Table 7) reveals the intrinsic structure of "cognitive space" in the residual stream. This space contains at least 3–4 distinguishable independent directions:

1. **Professional depth direction** (expert/guru/debugger cluster, pairwise 0.54–0.65)
2. **Pedagogical guidance direction** (socratic, nearly orthogonal to all)
3. **Critical analysis direction** (critic, only 0.19 with socratic)
4. **Simplification direction** (novice/child, weakly correlated with the professional depth direction)

This finding complements Anthropic's Assistant Axis work (arXiv:2601.10387), which identified an "assistant axis" in language models—a direction defining the default persona. Our results demonstrate that **the persona space extends far beyond a single axis**: different types of cognitive operations (deep analysis, guided questioning, critical examination, simplified explanation) each occupy independent directions. The "assistant" is a specific coordinate within this multi-dimensional space, not the space itself.

6.4 Manifold Interpretation of Steering Overshoot

At $\lambda > 1.5$, steering exhibits pronounced overshoot: SAF inflates sharply from ~100 to over ~1000, while cosine similarity actually decreases. This nonlinear behavior suggests that representations in the residual stream are not linearly distributed—they reside on a **curved low-dimensional manifold**.

Taking a small step along the steering direction ($\lambda \leq 1.0$) approximates curved motion along the manifold surface, successfully approaching the target persona. However, stepping too far ($\lambda > 2$) departs from the manifold surface into regions of the high-dimensional space with no corresponding semantic content—the model is pushed beyond its training distribution, causing massive spurious feature activation.

This sets a practical validity bound for steering methods: **a steering vector is a tangent approximation to the manifold, and its effectiveness is limited by the local curvature of the manifold**.

6.5 Unexplained Variance and Multi-Dimensional Superposition

The first principal direction for each persona explains 66–82% of the variance (Table 4). **What accounts for the remaining 18–34%?**

A single direction can capture the "magnitude" of steering—how far activations are pushed along a given direction. What it cannot capture is the "character" of steering—**to which region of the manifold** activations are pushed. The qualitative differences among personas—for instance, why expert covers higher dimensionality with fewer features—reside in those residual components.

This raises a conjecture worth pursuing: if standard persona prompts operate primarily along a single direction (one-dimensional steering), **might there exist prompting strategies that simultaneously produce significant projection components along multiple orthogonal directions?** Such prompts—if they exist—would produce effects irreducible to any linear combination of individual personas.

From a geometric perspective, a single persona is a ray in cognitive space (one direction). A naive concatenation of multiple personas yields a piecewise combination of such rays. But if a prompting strategy could **simultaneously activate multiple orthogonal steering components** in the residual stream, the resulting representation would occupy a high-dimensional subspace—beyond what any single ray can cover. This may explain why certain carefully crafted compound prompting strategies produce effects that far exceed the simple superposition of their individual components.

Verifying this conjecture requires systematically analyzing how complex prompt activations project onto the known persona steering directions. We leave this as a direction for future work.

6.6 Limitations

1. **Layer 50 only:** Goodfire's SAE is trained on Layer 50; Zhao (2026) shows EID peaks near Layer 70. The most critical feature differentiation may occur at deeper layers.
2. **Feature labels are inferred:** We infer semantics from activation patterns but lack direct semantic validation (e.g., Neuronpedia labels). AutoInterp's semantic labels are hypotheses, not proofs.
3. **Correlation, not causation:** We demonstrate the existence of feature differences and steering effects, but have not established causation through intervention experiments (amplifying/suppressing specific features and observing output changes).
4. **Single model:** All results are from Llama-3.3-70B-Instruct. Whether the cognitive dimension space structure generalizes across models requires further investigation.
5. **Linear steering assumption:** We employ linear steering (additive superposition), but the true geometry of the residual stream may be nonlinear. The overshoot phenomenon (Section 6.4) already hints at this limitation.
6. **Chinese-language prompts:** The experiments use Chinese-language prompts. Whether different languages produce identical steering geometries remains an open question.

7. Conclusion

This paper systematically analyzes the representational changes induced by prompts within language models through three experimental tiers: the pilot study (6 conditions × 50 topics), the persona experiment (16 conditions × 100 topics), and the steering experiment. The principal findings are:

SAF and EID are orthogonal cognitive measures: - SAF reflects cognitive complexity (how many semantic units are recruited), while EID reflects cognitive dimensionality (how large a semantic space is covered). The two can vary independently—expert covers the highest dimensionality (EID rank 3/16) with the fewest features (SAF rank 13/16), while socratic activates the most features (SAF rank 1/16) but achieves only moderate dimensionality (EID rank 6/16).

Role assignment is dimensional expansion: - All 14 non-baseline personas yield nEID between 1.52 and 2.18, indicating that merely assigning the model a role identity is sufficient to elevate representational dimensionality by over 50%. Debugger achieves the highest EID (nEID 2.18), as debugging scenarios demand the broadest semantic space coverage.

Persona prompts approximate one-dimensional steering: - In the 8,192-dimensional residual stream, persona-induced displacements are 66–82% explainable by a single direction. At $\lambda = 1.0$, cosine similarity between steered standard and true persona exceeds 0.90 for all cases, and Pearson correlation exceeds 0.90 throughout.

Cognitive space is a multi-dimensional manifold: - The steering directions of different personas form a multi-dimensional "cognitive dimension space." The expert–guru–debugger cluster shares a professional depth direction (pairwise cosine 0.54–0.65), socratic exclusively occupies a pedagogical guidance direction (nearly orthogonal to all others), and critic exclusively occupies a critical analysis direction. The cosine between novice and expert is only 0.46—the novice is not an inverse expert but a distinct cognitive dimension.

These results decompose Zhao (2026)'s global EID findings to the level of interpretable features and geometric structure. The "cognitive space" within language models possesses a rich multi-dimensional structure, with different prompts pushing representations along different directions whose effects are mutually irreducible. Understanding this structure will provide a foundation for more principled prompt engineering and more fine-grained model interpretability research.

References

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

Chanin, D., Wilken-Smith, B., Dulka, T., Bhatnagar, A., & Bloom, J. (2026). The Assistant Axis: Situating and Stabilizing the Default Persona of Language Models. *arXiv:2601.10387*.

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

Li, H., Yang, X., Yao, Z., et al. (2025). Feature Extraction and Steering for Enhanced Chain-of-Thought Reasoning in Language Models. *arXiv:2505.15634*.

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
- **Pilot study data:** `feature_diff.json`, `feature_context.json`, `autointerp_results.json` in repository
- **Persona experiment data:** `activations_persona_layer50.pt`, `features_persona_layer50.pt`, `eid_persona_results.json`
- **Steering experiment data:** `steering_analysis_results.json`
- **UMAP visualizations:** `umap_activations.png`, `umap_features.png` in repository

Version History: - v1 (2026-01-31): Initial release with activation count and exclusive feature analysis - v2 (2026-02-01): Added AutoInterp feature semantic analysis, revealing semantic subdivision structure - v3 (2026-02-02): Added UMAP visualization, showing spatial separation and SAE denoising effect - v4 (2026-02-15): Major upgrade—added Persona experiment (16 conditions × 100 topics), EID analysis, Steering experiment, revealing multi-dimensional manifold structure of the cognitive dimension space

Last updated: 2026-02-15