Invariant Features in Language Models: Geometric Characterization and Model Attribution

Agnibh Dasgupta, Abdullah All Tanvir, Xin Zhong | University of Nebraska Omaha | NeurIPS 2026

Background: LLMs are robust to paraphrasing.



"A man is riding a horse."

"He is riding a horse."

"A person is on horseback."

Different wording → same meaning

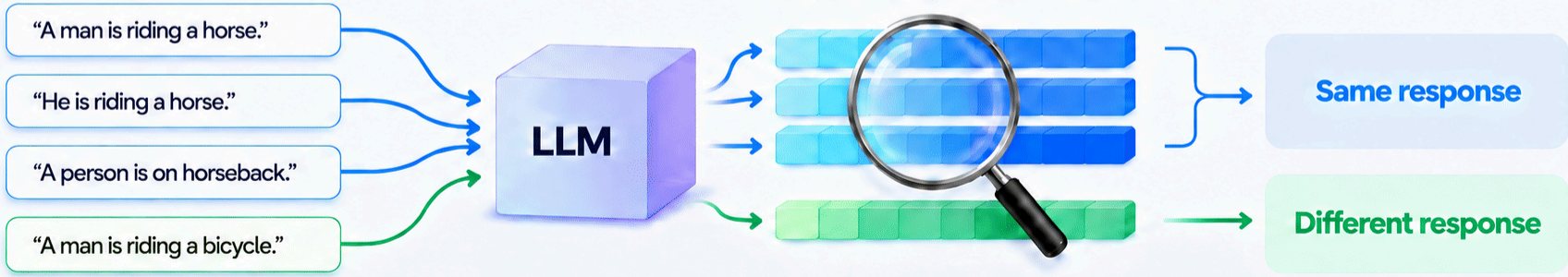


"A man is riding a horse."



"A man is riding a bicycle."

Different meaning → semantic change

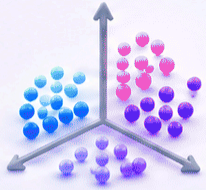


Prior Work



Meaning in Model Layers

- Lower layers encode syntax, whereas higher layers capture context and semantics.
  (Tenney et al. 2019; Hewitt & Manning 2019).
- Paraphrased inputs yield closely aligned internal representations across layers.
  (Ethayarajh 2019; Reif et al. 2019).



Representation Geometry

- Metrics like CCA, SVCCA, PWCCA, and CKA measure representational similarity across layers and networks.
  (Raghu et al. 2017; Kornblith et al. 2019).
- Linear probes isolate low-dimensional subspaces dedicated to specific linguistic features.



Model Identity

- Soft prompts and adversarial triggers exhibit model-specific behaviors across architectures.
  (Lester et al. 2021; Wallace et al. 2019).
- Current model attribution relies on generated outputs, watermarking, or training-phase markers.
  (Antoun et al. 2024; Wang et al. 2023).



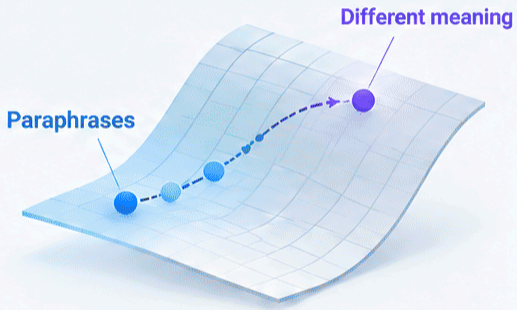
Our Research Question

Can we isolate directions inside the model that remain stable under paraphrasing yet respond to meaning shifts, and do these directions distinguish individual models?

Our Contributions

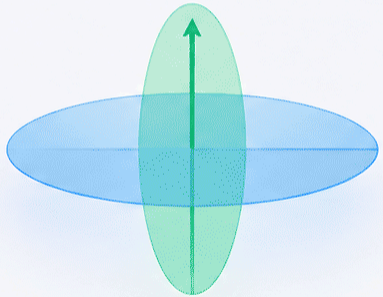
A local geometric framework that separates paraphrase variation from meaning, and uses it to identify models.

1 Geometric Characterization



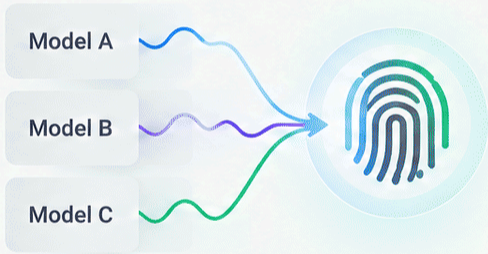
We define invariant features via a local manifold view separating paraphrases from meaning.

2 Contrastive Subspace Discovery



A training-free generalized eigenvalue method isolating directions invariant to paraphrasing.

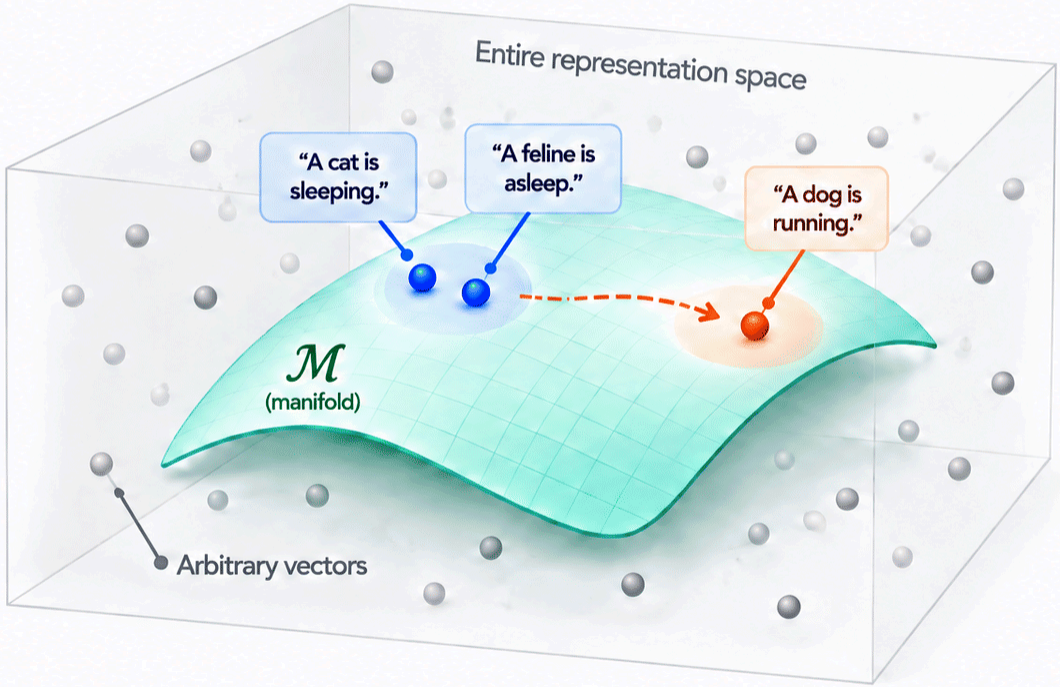
3 Zero-Shot Model Attribution



Invariant subspaces serve as unique model-specific signatures without classifier training.

Hypothesis: Low-Dimensional Manifold

Manifold hypothesis: high-dimensional data concentrate near a much lower-dimensional manifold (Fefferman et al., 2016).



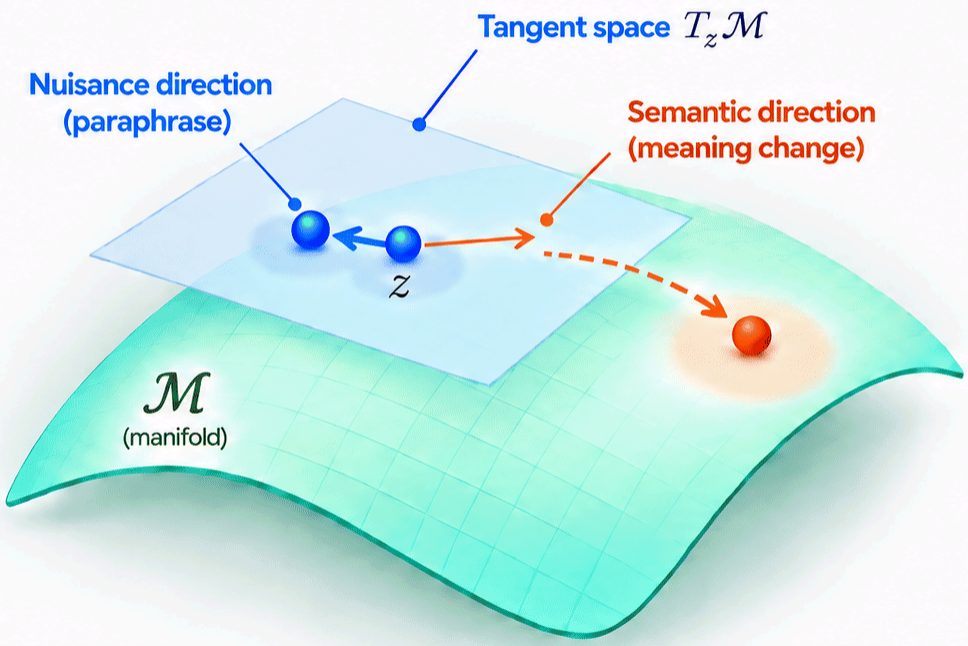
$\mathcal{M}$: meaningful representations lie near a low-dimensional surface.

Within a region: wording, syntax, style (same meaning).

Across regions: meaning changes.

We apply this to LLM hidden states: representations of meaningful inputs lie near a manifold. $\mathcal{M}$.

Tangent Space Decomposition



1 **Local approximation:** near z, the manifold is approximately flat, and its tangent space splits into two kinds of directions.

$$T_z\mathcal{M} \approx T_z^{\text{semantic}} \oplus T_z^{\text{nuisance}}$$

2 **Nuisance motion:** paraphrasing moves the representation mainly along nuisance directions.

$$z' = z + v_{\text{nuisance}}$$

3 **Projection to the semantic subspace:** projecting out nuisance directions gives an invariant feature.

$$z_{\text{inv}} = Pz, \quad \text{so} \quad P(z + v_{\text{nuisance}}) = Pz$$

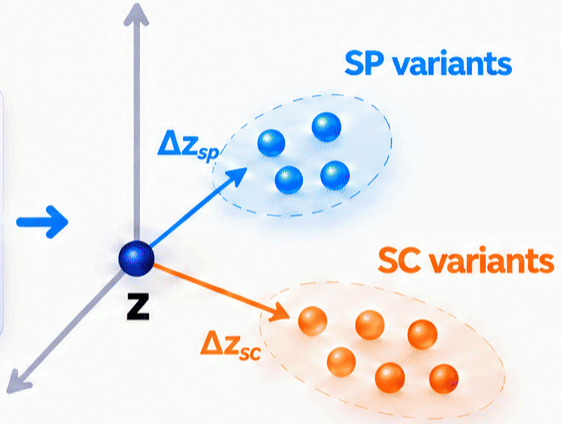
Local and approximate. Semantic and nuisance directions may partly overlap.

Contrastive Data: SP vs SC Variants

Anchor sentence

"What is the official language of France?"

LLM



Variant types and examples

Anchor	"What is the official language of France?"
SP	"Which language holds official status in France?"
SC	"What is the national currency of France?"

1,400 semantic groups

13 topics

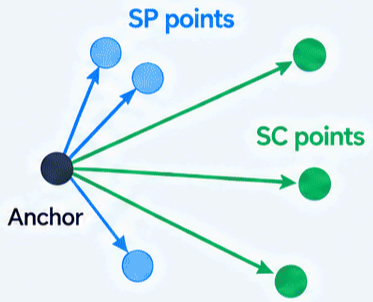
57,400 sentences

Per topic:
1 anchor, 20 SP,
10 soft SC, 10 hard SC

Find directions where SC variation is large relative to SP variation.

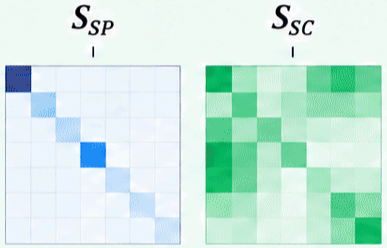
Contrastive Sensitivity Decomposition

1 Displacements



Change in representation relative to the anchor.

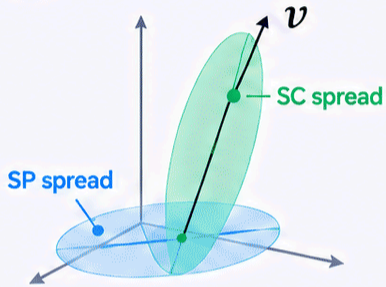
2 Covariance matrices



Illustrative

How SP and SC edits move the representation.

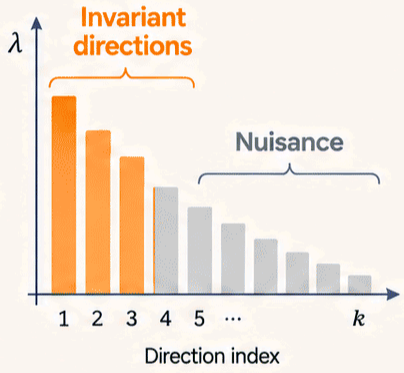
3 Generalized eigenvalue problem



$$\lambda(v) = \frac{v^{\mathrm{T}} S_{SC} v}{v^{\mathrm{T}} S_{SP} v}$$

v: a direction in representation space.
λ is high when SC variation is large relative to SP variation (Rayleigh quotient).

4 Top-k directions

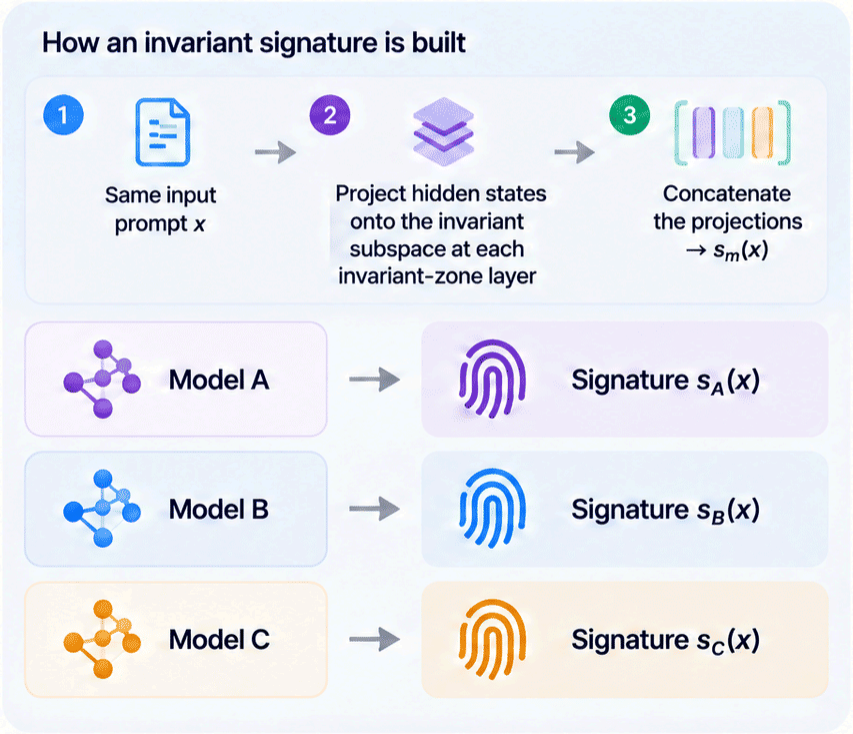


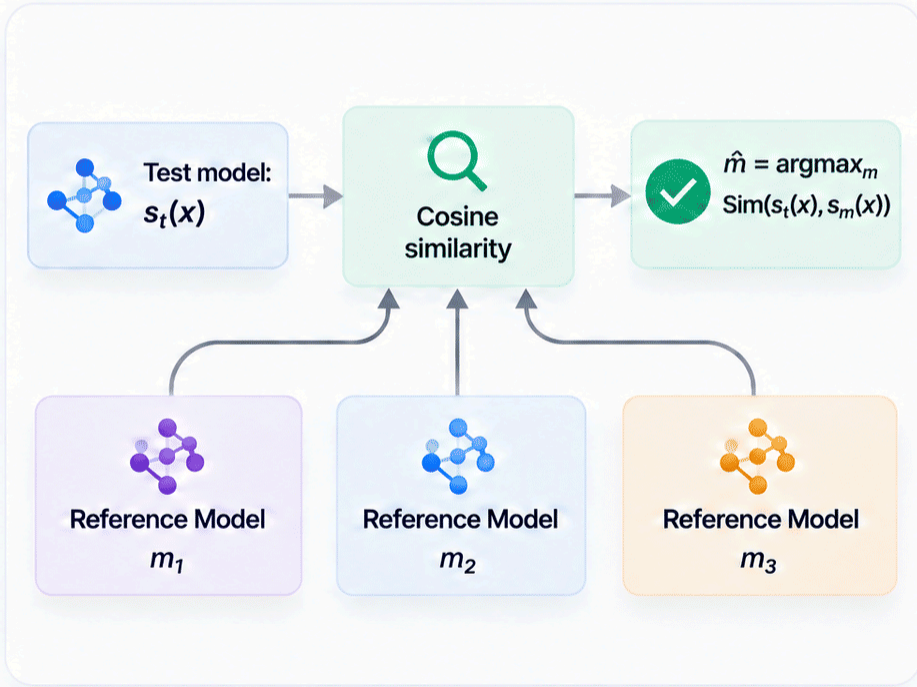
The top-k eigenvectors span the invariant subspace.

Computed at every layer. Layers with the largest normalized λ form the invariant zone.

S_{SP} is low rank, so we solve with $S_{SP} + \varepsilon I$.

Zero-Shot Model Attribution with Invariant Signatures





Experimental Setup

Semantic Groups

Examples from 1,400 groups across 13 domains.

Group	SP	Soft SC	Hard SC
Cats	20	10	10
France	20	10	10
Python	20	10	10
Photosynthesis	20	10	10
Climate change	20	10	10
...	...	...	...

Evaluated Models

Nine models across families, scales (4B to 16B) and training setups.

Model	Family	Parameters
Mistral-Instruct-v0.3	Mistral	7B
Llama-3-Instruct	Llama	8B
Gemma-IT	Gemma	7B
Qwen2.5-Instruct	Qwen	7B
GLM-4-Chat	GLM	9B
DeepSeek-MoE-Chat	DeepSeek	16B
InternLM2-Chat	InternLM	7B
Phi-4-Mini	Phi	4B
Falcon3-Instruct	Falcon	7B

Example: One Prompt, Two Variants

Anchor sentence (original)

"Cats are mammals that purr."

→

SP: Meaning-preserving paraphrase

"Cats are mammals that make a purring sound."

→

SC: Meaning-changing variant

"Cats are reptiles that purr."

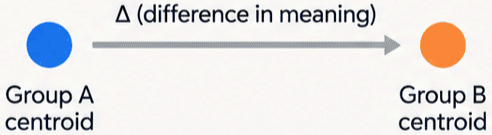
Results: Nuisance Projection Energy

1 Split each hidden state

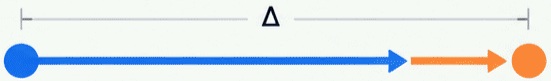


Each hidden state is split by projection onto the invariant subspace and its complement.

2 Meaning difference

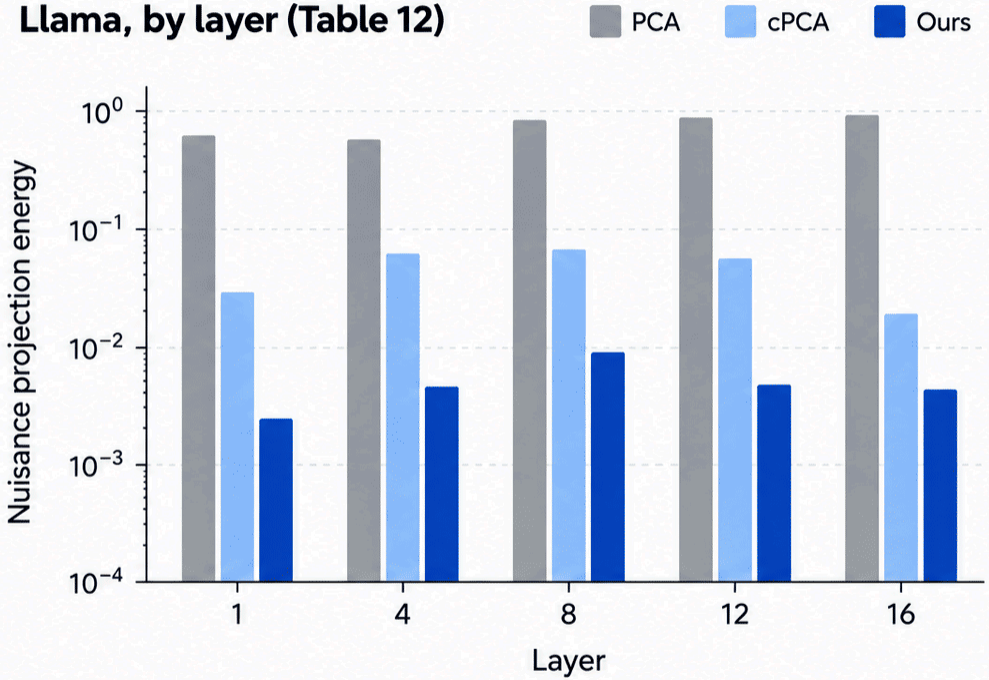


3 Energy in the nuisance part



Nuisance projection energy = share of Δ that falls in the nuisance part (0 to 1).

Llama, by layer (Table 12)



Displacement between groups puts little projection energy in the nuisance subspace, far less than PCA or cPCA.

Results: Causal Validation with Interventions

1. Original and donor inputs



What do cats usually eat?



What do dogs usually eat?

2. Representations (layers 1 to ℓ)

LLM (layers 1 to ℓ)

Same LLM (layers 1 to ℓ)

Original hidden state

Invariant part (inv) | Nuisance part (nuis)

Donor hidden state

Invariant part (inv') | Nuisance part (nuis')

3. Patch (copy donor's invariant part)

Patched hidden state

Invariant part (inv') | Nuisance part (nuis)

4. Continue (layers ℓ+1 to L)

Same model continues (layers ℓ+1 to L)

5. Output

Original output:

Cats usually eat meat, such as ...

Patched output:

Dogs usually eat ...

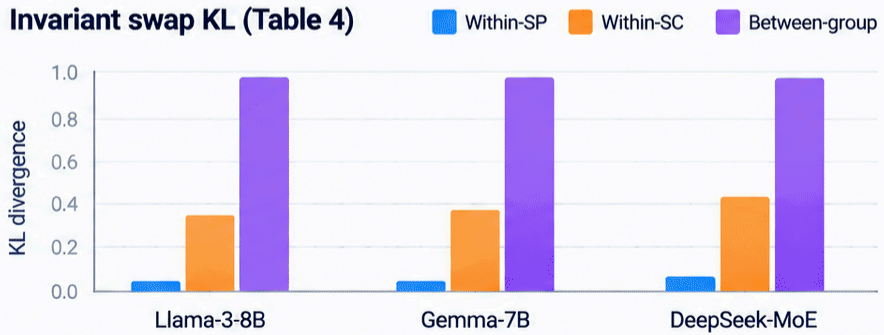
SP donor: same meaning

"What food do cats typically eat?" → **Output changes little.**

SC donor: different meaning

"What do dogs usually eat?" → **Output changes substantially.**

KL divergence measures how much the output distribution changes.



Final-token hidden state, one representative group, averaged over the top 4 invariant layers.

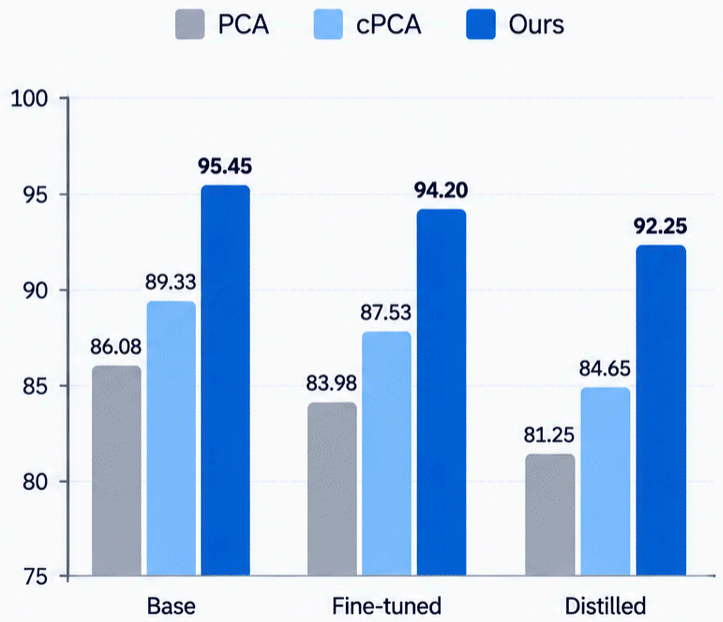
Results: Zero-Shot Model Attribution

Attribution accuracy % (Table 5)

Model	Base	Fine-tuned	Distilled
Mistral-7B	93.2	91.4	89.8
Llama-3-8B	94.1	92.5	90.9
Gemma-7B	92.8	90.9	89.2
Qwen2.5-7B	93.6	91.8	90.1
Phi-4-mini	89.6	87.8	86.1
GLM-4-9B	95.2	93.6	91.9
DeepSeek-MoE	95.8	94.2	92.6
InternLM2-7B	94.9	93.3	91.7
Falcon3-7B	95.5	93.9	92.2

Group 1. Signatures come from base models only and are reused unchanged for the fine-tuned (LoRA on Alpaca) and distilled variants.

Baselines, averaged over 4 models (Table 6)



Limitations and Future Work

Limitations

1

Local, not universal

Each semantic group has its own invariant subspace. The decomposition is local and approximate, not a single subspace for all meaning.

2

Controlled data

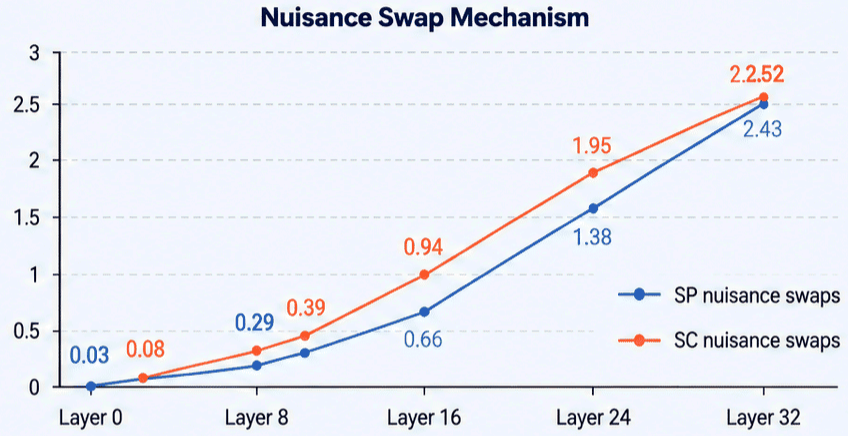
The dataset is synthetic, English and short. Naturally occurring, multilingual and long-context inputs remain to be evaluated.

3

Partial disentanglement

Nuisance swaps still change outputs, even with paraphrase donors, so the nuisance part may retain some semantic information.

Evidence for Partial Disentanglement



SP and SC nuisance swaps rise together with depth, without consistent separation (Fig. 5).

Future Work

Subspace Refinement
More complete disentanglement of the nuisance subspace.

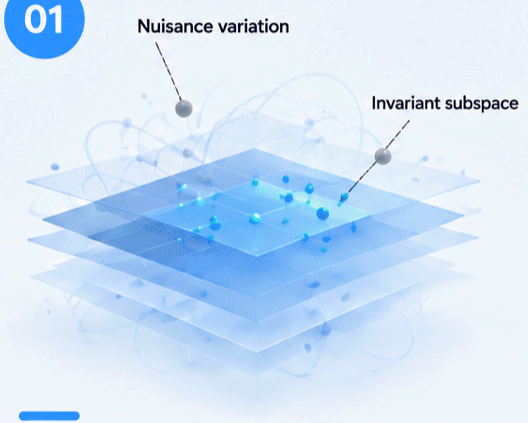
Broader Evaluation
Evaluation on natural, multilingual and long-context inputs.

Cross-Layer Relations
Relating invariant subspaces across layers and across semantic groups.

Conclusion & Key Takeaways

A local geometric framework for identifying semantic invariance in LLM representations, and using it to attribute models.

01

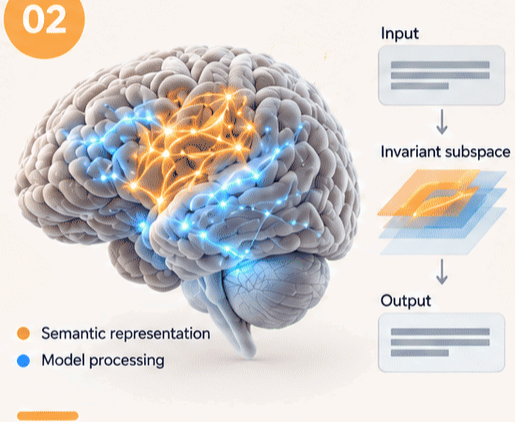


A Geometric Framework

Meaning differences fall largely outside the nuisance subspace.

Invariant structure is strongest in mid-to-late layers.

02



Evidence of a Causal Role

Swapping the invariant part changes outputs when meaning changes, and barely when it does not.

03



A Robust Fingerprint

Invariant signatures identify models zero-shot, including fine-tuned and distilled variants, and outperform PCA and cPCA.