# Invariant Features in Language Models: Geometric Characterization and Model Attribution

**Agnibh Dasgupta*** **Abdullah All Tanvir*** **Xin Zhong†**
Department of Computer Science
University of Nebraska Omaha, Omaha, NE, USA

## Abstract

Language models exhibit strong robustness to paraphrasing, suggesting that semantic information may be encoded through stable internal representations, yet the structure and origin of such invariance remain unclear. We propose a local geometric framework in which semantically equivalent inputs produce representations that occupy structured regions in latent space, with paraphrastic variation along nuisance directions and semantic identity preserved in invariant subspaces. Building on this view, we make three contributions: (1) a geometric characterization of invariant latent features, (2) a contrastive subspace discovery method that separates semantic-changing from semantic-preserving variation, and (3) an application of invariant representations to zero-shot model attribution. Across models and layers, empirical results support these contributions. Invariant structure emerges in specific depth regions, semantic displacement lies largely outside the nuisance subspace, and representation-level interventions indicate a causal role of invariant components in model outputs. Invariant representations also capture model-specific geometric patterns, enabling accurate attribution. Together, these findings support a local geometric characterization of semantic invariance in language model representations.

## 1 Introduction

Large language models (LLMs) exhibit a striking robustness to surface-level variation. Prompts that differ substantially in wording, style, or syntactic structure often lead to similar internal reasoning and consistent outputs. This behavior suggests that LLMs encode semantic information in a stable internal representation that abstracts away from lexical form. However, despite extensive empirical success, it remains unclear where such stable semantic representations emerge within model architectures, how they can be formally characterized, and how their geometric structure differs across models.

Understanding invariant internal representations is important for both scientific and practical reasons. From a representation learning perspective, identifying where semantic stability arises provides insight into how transformers organize meaning across depth and how abstraction emerges during inference. From a systems and security perspective, invariant representations offer a mechanism to analyze model behavior beyond observable outputs. Because these internal structures reflect how a model organizes semantic information, they may support applications such as model attribution, forensic analysis, and robustness evaluation.

Existing efforts to analyze or identify language models span several complementary directions. Prior work on semantic representations and probing has shown that transformer activations encode structured linguistic and semantic information and exhibit robustness to paraphrasing. A separate line

*Equal contribution.
†Corresponding author: xzhong@unomaha.edu

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

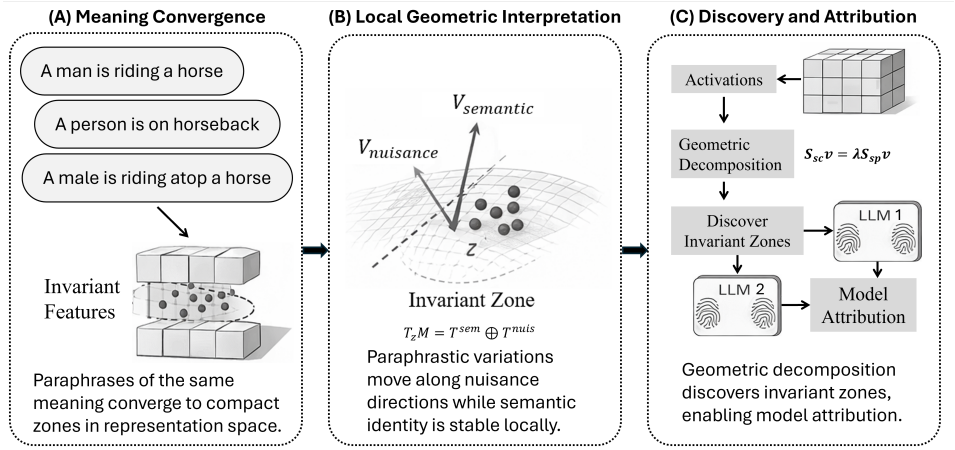


Figure 1: Emergence of invariant semantic regions in LLMs. (A) Paraphrases propagate through layers and tend to converge to compact zones. (B) Local geometric interpretation: nuisance vs semantic tangent directions. (C) Geometric decomposition detects invariant zones for model attribution.

of work studies representation geometry through similarity metrics and subspace analysis, revealing that learned features admit meaningful low-dimensional structure across layers and models. More recently, model attribution and fingerprinting approaches have demonstrated that internal activations and outputs can contain model-specific signals. While these studies provide important insights, they largely treat semantic stability, representation geometry, and model identity as separate phenomena. In particular, existing approaches do not provide a local geometric characterization of semantic invariance as a structural property of latent representations, nor a principled mechanism to identify invariant structure by explicitly separating semantic-preserving and semantic-changing variation. As a result, it remains difficult to systematically understand how semantic invariance emerges across layers and how it relates to model-specific representation geometry.

In this work, we develop a geometric framework for analyzing semantic invariance in language models. As illustrated in Fig. 1, we hypothesize that semantically equivalent inputs form locally stable regions in latent space, where semantic-preserving and semantic-changing variations correspond to approximately distinct nuisance and semantic directions. (1) Building on this hypothesis, our first contribution is a local geometric definition of invariant latent features as representations that are stable under semantic-preserving transformations while remaining discriminative across meanings. (2) Our second contribution is a data-driven framework for identifying local invariant features via contrastive sensitivity decomposition, which separates semantic-changing from semantic-preserving variation. (3) Our third contribution is to show that these invariant zones exhibit model-specific geometry and can serve as latent signatures for zero-shot model attribution, demonstrating that invariant representations encode both semantic stability and model individuality.

## 2 Related Work

**Semantic Representations and Emergent Invariance in Language Models.** Prior work shows that transformer-based language models develop structured representations across depth, with lower layers encoding lexical and syntactic information and higher layers capturing contextual and semantic properties [1, 2]. Probing studies further demonstrate that linguistic structures such as syntax and dependency relations can be recovered from these representations [3, 4]. Geometric analyses of BERT additionally suggest that semantic and syntactic information occupy structured subspaces in latent space, with word senses exhibiting fine-grained organization [5]. Together, these findings indicate that language models encode semantic structure beyond surface form. This line of work also implies an empirical form of semantic invariance: inputs that differ in wording but preserve meaning often yield similar representations. Analyses of contextualized embeddings and their reductions to static representations further show that semantic information remains stable across contexts [1, 6].

Existing studies primarily establish that semantic structure is present in language-model representations. However, a local geometric characterization of how semantic invariance emerges across layers, how paraphrastic variation separates from semantic variation, and how such invariant structure can be operationally identified, remains less explored.

**Representation Geometry and Subspace Analysis in Neural Networks.** A central line of work studies neural representations through geometric and statistical comparisons of activation spaces. Representational similarity analysis (RSA) and related methods quantify relationships across inputs, layers, and models. Widely used techniques include canonical correlation analysis (CCA) and its variants such as singular vector CCA (SVCCA) [7] and projection-weighted CCA (PWCCA) [8], as well as centered kernel alignment (CKA), which is invariant to orthogonal transformations and isotropic scaling [9]. These methods enable systematic comparison of representations and are widely used to analyze training dynamics, layer correspondence, and cross-model similarity. Complementary approaches examine the structure of representation spaces directly. Linear probes show that linguistic and semantic features can be recovered from learned representations, while geometric analyses indicate that such information resides in structured subspaces [3, 5]. Together, these results suggest that neural representations admit meaningful low-dimensional structure that can be characterized through linear projections and subspace decompositions.

Despite these advances, existing approaches mainly measure similarity or identify directions associated with specific properties. They do not explicitly provide mechanisms for separating semantic-preserving from semantic-changing variation. Consequently, while prior work reveals structure in representation space, a principled framework for isolating invariant subspaces that distinguish semantic content from nuisance variation remains limited. This motivates our formulation of invariant subspace discovery as a contrastive geometric decomposition, where invariant directions arise from a generalized eigenvalue operator that separates semantic and nuisance components.

**Model-Specific Representations and Attribution via Internal Activations.** A growing body of work suggests that language models encode model-specific structure despite similar external behavior. Evidence arises from transfer studies, where soft prompts optimized for one model often fail to generalize to others, indicating interaction with model-specific latent pathways [10]. Related observations come from adversarial trigger analysis, where input-agnostic triggers reveal global biases and model-dependent vulnerabilities [11]. These results suggest that similar linguistic functions may be implemented through different internal geometries. Model attribution has also been studied via generated text, watermarking, and explicit identification signals. Prior work explores source-model identification from outputs [12], watermark-based attribution [13, 14], and instruction-based identification through lightweight training interventions [15].

While these approaches demonstrate that model identity can be detected, they typically rely on output-level artifacts, task-specific classifiers, watermarking mechanisms, or training-induced signatures. In contrast, our work uses invariant latent structure itself as a compact semantic signature. Attribution is therefore not treated as a standalone identification mechanism, but as a downstream validation that invariant subspaces preserve both semantic stability and model-specific representation geometry.

## 3 Proposed Framework for Invariant Latent Feature Analysis

We present a geometric framework for semantic invariance in language models, based on the hypothesis that paraphrastic variation induces structured zones of stability. Section 3.1 formalizes invariant latent features, Section 3.2 introduces a contrastive subspace discovery method, and Section 3.3 uses the resulting representations for zero-shot model attribution as empirical validation.

### 3.1 Invariant Latent Features: Definition and Local Geometry

Let $f_m$ denote a language model, $h_\ell(x) \in \mathbb{R}^{T\times d}$ denote the token-level hidden states at layer $\ell$, and $z = \Phi_\ell(x) \in \mathbb{R}^d$ denote their pooled representation, which serves as the primary object of analysis. We study invariance of these representations under semantic-preserving transformations. Formally, for a transformation $T(\cdot)$ such as paraphrasing, invariance requires $\Phi_\ell(T(x)) \approx \Phi_\ell(x)$. To avoid trivial collapse, representations must also remain discriminative across meanings. Specifically, for paraphrases $x_1, x_2$ and an input $x_3$ with different meaning, we require $\Phi_\ell(x_1) \approx \Phi_\ell(x_2), \quad \Phi_\ell(x_1) \not\approx \Phi_\ell(x_3)$. This within–between criterion defines invariant latent features as representations that are stable under semantic-preserving variation while remaining separable across semantic content.

To interpret this property, we adopt a geometric perspective inspired by the manifold hypothesis [16]. Although latent representations reside in a high-dimensional ambient space $\mathbb{R}^d$, meaningful semantic information occupies only a structured subset of this space. Let $\mathcal{M} \subset \mathbb{R}^d$ denote the semantic

representation manifold. Inputs with coherent semantic meaning are mapped to points lying near this manifold, while arbitrary points in the ambient space generally do not correspond to meaningful language representations. Under this view, semantically equivalent inputs correspond to nearby points on the manifold. Paraphrasing can therefore be interpreted as a local perturbation that moves the representation within a neighborhood of the same semantic region. Specifically, for paraphrases $x_1$ and $x_2$, $\|\Phi_\ell(x_1) - \Phi_\ell(x_2)\| \leq \epsilon$, for some small $\epsilon > 0$, while input $x_3$ with distinct semantic meaning occupy separate regions of the manifold.

We propose to analyze the local structure of $\mathcal{M}$ through its tangent space. At a point $z = \Phi_\ell(x)$ on the manifold, the tangent space $T_z\mathcal{M}$ contains all directions along which the representation can move infinitesimally while remaining on the manifold. In other words, $T_z\mathcal{M}$ provides a local linear approximation of the manifold and captures the allowable directions of variation in representation space. Figure 2 illustrates this local geometric interpretation. Variations of representations on the manifold arise from two qualitatively different sources: semantic changes, which alter the underlying meaning of the input, and nuisance changes, which modify surface form while preserving semantic content. Locally, these two types of variation can be interpreted as approximately distinct directions within the tangent space. We therefore consider a working decomposition of the form

$$T_z\mathcal{M} \approx T_z^{\text{semantic}} \cup T_z^{\text{nuisance}} \tag{1}$$

where $T_z^{\text{semantic}}$ captures directions associated with changes in semantic content, and $T_z^{\text{nuisance}}$ captures directions corresponding to semantic-preserving variation. In practice, this decomposition is a local and approximate factorization, since learned representations may exhibit partial overlap and entanglement between these components. Under a semantic-preserving transformation such as paraphrasing, the representation primarily moves along nuisance directions, $z' = z + v_{\text{nuisance}}$. Invariant features correspond to representations that ignore such nuisance motion. If $P$ denotes a projection operator onto the semantic subspace, the invariant representation can be written as $z_{\text{inv}} = Pz$, which satisfies $P(z + v_{\text{nuisance}}) = Pz$. Throughout this work, "invariant" refers specifically to invariance with respect to semantic-preserving transformations. Thus, invariant features are expected to remain stable under paraphrasing while remaining sensitive to semantic-changing perturbations.

This local geometric interpretation provides an explanation for why invariant latent features naturally emerge during training. To maintain consistent behavior under paraphrasing, the model must map semantically equivalent inputs to nearby regions in latent space while preserving separability across meanings. These constraints induce structured zones of stability on the representation manifold where paraphrase variants cluster while remaining distinguishable from other semantic concepts.

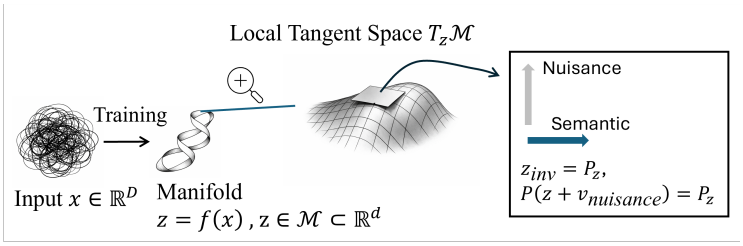


Figure 2: Local geometric hypothesis of semantic invariance in LM representations. During training, inputs $x \in \mathbb{R}^D$ are mapped by the model to latent representations $z = f(x)$ that lie on a lower-dimensional manifold $\mathcal{M} \subset \mathbb{R}^d$. At a point $z \in \mathcal{M}$, the tangent space $T_z\mathcal{M}$ provides a local linear approximation of allowable variations in representation space. Semantic-preserving transformations such as paraphrasing primarily induce movements along nuisance directions that do not alter the underlying meaning. Invariant latent features correspond to projections that remove these nuisance directions, so that $z_{\text{inv}} = P_z$ and $P(z + v_{\text{nuisance}}) = P_z$, preserving semantic identity while ignoring surface variation.

Although the existence of such invariant zones follows from the need for semantic consistency, their geometric realization is generally model-specific. Transformer training involves highly non-convex optimization [17, 18], and different models may converge to distinct embeddings of the underlying semantic manifold due to variations in initialization, architecture, training data, and optimization dynamics. Consequently, even when two models encode similar semantic concepts, the corresponding latent manifolds may differ in orientation, curvature, and local geometry. Invariant latent features therefore reflect how a particular model internally organizes semantic information. This combination of semantic stability and geometric individuality later enables these invariant representations to serve as distinctive identifiers for model attribution.

### 3.2 Invariant Subspace Identification via Contrastive Sensitivity Decomposition

The geometric formulation in Section 3.1 characterizes invariant latent features as directions that remain stable under semantic-preserving transformations while remaining responsive to semantic change. We now introduce a data-driven procedure to identify such directions from model activations. The key idea is to contrast two sources of variation in latent space: variation induced by semantic-preserving transformations, which primarily reflects nuisance motion, and variation induced by semantic-changing transformations, which reflects movement across meanings.

Let $x$ denote an input prompt, at layer $\ell$, let $h_\ell(x) \in \mathbb{R}^{T\times d}$ denote token-level hidden states. We define a representation vector $z_\ell(x) = \Phi_\ell(x) \in \mathbb{R}^d$ as a deterministic pooling of $h_\ell(x)$. In practice, we use the final-token hidden state, i.e., $z_\ell(x) = h_\ell^{(T)}(x)$, which incorporates information from preceding tokens and directly supports next-token prediction. We compare this choice with mean pooling in Appendix H. For each input, we construct two perturbation families. The first consists of semantic-preserving variants, denoted by $\{x_i^{\mathrm{sp}}\}_{i=1}^n$, such as paraphrases that preserve the underlying meaning while altering surface form. The second consists of semantic-changing variants, denoted by $\{x_j^{\mathrm{sc}}\}_{j=1}^m$, whose semantic content differs from that of the anchor input. We then compute the corresponding finite-difference feature changes $\Delta z_i^{\mathrm{sp}} = \Phi_\ell(x_i^{\mathrm{sp}}) - \Phi_\ell(x)$, $\Delta z_j^{\mathrm{sc}} = \Phi_\ell(x_j^{\mathrm{sc}}) - \Phi_\ell(x)$. These difference vectors provide empirical estimates of local motion in the representation space. Under the local geometric hypothesis, semantic-preserving perturbations primarily trace nuisance directions in the tangent space, whereas semantic-changing perturbations additionally traverse semantic directions. To capture these two modes of variation, we define the empirical sensitivity covariances $S_\ell^{\mathrm{sp}} = \frac{1}{n}\sum_{i=1}^n \Delta z_i^{\mathrm{sp}}(\Delta z_i^{\mathrm{sp}})^\top$, $S_\ell^{\mathrm{sc}} = \frac{1}{m}\sum_{j=1}^m \Delta z_j^{\mathrm{sc}}(\Delta z_j^{\mathrm{sc}})^\top$. We then identify contrastive directions by solving the generalized eigenvalue problem:

$$S_\ell^{\mathrm{sc}} v = \lambda\, S_\ell^{\mathrm{sp}} v. \tag{2}$$

Each eigenvector $v$ maximizes the Rayleigh quotient $\lambda(v) = v^\top S_\ell^{\mathrm{sc}} v / v^\top S_\ell^{\mathrm{sp}} v$. This formulation encodes the desired notion of the proposed invariance. Directions with large $\lambda$ exhibit strong variation under semantic change but weak variation under semantic-preserving perturbations, and therefore correspond to invariant semantic directions. In contrast, directions with small $\lambda$ are dominated by semantic-preserving variation and are interpreted as nuisance directions. We therefore use the term invariant direction to denote a direction that is stable under semantic-preserving perturbations but discriminative under semantic-changing perturbations. All covariance estimates and the generalized eigenvalue problem are computed on $z_\ell(x)$.

In practice, $S_{\mathrm{sp}}$ is low rank because the number of semantic-preserving perturbations is much smaller than the hidden dimension. We therefore solve a regularized generalized eigenvalue problem, which ensures positive definiteness of the denominator operator and yields a well-posed decomposition, $S_{\mathrm{sc}} v = \lambda(S_{\mathrm{sp}} + \epsilon I)v$, where $\epsilon > 0$ ensures numerical stability. Let $V_\ell^{\mathrm{inv}} = [v_1, \ldots, v_k]$ denote the top-$k$ generalized eigenvectors with the largest eigenvalues at layer $\ell$. Since generalized eigenvectors are orthogonal with respect to the $(S_{\mathrm{sp}} + \epsilon I)$-inner product but not necessarily under the Euclidean inner product, we compute an orthonormal basis $U_\ell^{\mathrm{inv}} = \mathrm{orth}(V_\ell^{\mathrm{inv}})$ via QR decomposition. This ensures that $U_\ell^{\mathrm{inv}} U_\ell^{\mathrm{inv}\top}$ defines a valid Euclidean projection operator. The invariant representation at layer $\ell$ is then obtained by projection,

$$z_{\mathrm{inv}} = U_\ell^{\mathrm{inv}} U_\ell^{\mathrm{inv}\top} z. \tag{3}$$

This projection suppresses nuisance-sensitive components while preserving directions that remain stable under paraphrasing yet discriminate across meanings. The complementary nuisance component is given by $z_{\mathrm{nuis}} = \left(I - U_\ell^{\mathrm{inv}} U_\ell^{\mathrm{inv}\top}\right) z$. Together they induce the decomposition $z = z_{\mathrm{inv}} + z_{\mathrm{nuis}}$, providing a constructive approximation of the semantic–nuisance decomposition in Eq. 1. In particular, $z_{\mathrm{inv}}$ captures components aligned with invariant semantic directions, while $z_{\mathrm{nuis}}$ captures variation dominated by nuisance-sensitive directions.

The decomposition in Eq. 3 also admits a local differential interpretation. When perturbations are infinitesimal, $\Delta z \approx J_{\Phi_\ell}(x)\Delta x$, where $J_{\Phi_\ell}(x)$ is the Jacobian of the layer-wise representation map. The resulting covariance operators satisfy $S \approx J_{\Phi_\ell}(x)\,\mathbb{E}[\Delta x \Delta x^\top]\, J_{\Phi_\ell}(x)^\top$. Hence, the proposed covariance-based formulation can be viewed as a data-driven approximation of Jacobian sensitivity. Moreover, if the input perturbations are isotropic so that $\mathbb{E}[\Delta x \Delta x^\top] = I$, then $S \approx U\Sigma^2 U^\top$, where $J_{\Phi_\ell}(x) = U\Sigma V^\top$ is the singular value decomposition of the Jacobian. In that case, the

eigenvectors of $S$ align with the left singular vectors of the Jacobian, and the eigenvalues correspond to squared singular values. The generalized eigenvalue formulation therefore extends standard sensitivity analysis by explicitly separating semantic and nuisance variation.

We apply this procedure across layers and select those exhibiting strong invariant structure based on their generalized eigenvalue spectra. Invariant zones are thus defined as sets of layers associated with leading eigenvalues, without requiring spatial contiguity. For each model $m$ and input $x$, we define the invariant zone representation by aggregating invariant projections across the selected layers, $s_m(x) = \mathrm{Agg}\big(z_{\mathrm{inv}}^{(\ell_i)}, \ldots, z_{\mathrm{inv}}^{(\ell_j)}\big)$, where $\{\ell_i, \ldots, \ell_j\}$ denotes the set of layers whose invariant subspaces are identified by large generalized eigenvalues, and $\mathrm{Agg}(\cdot)$ denotes feature aggregation across layers, implemented as concatenation of the invariant representations. This representation serves as a compact semantic signature capturing the invariant structure of the model.

### 3.3 Zero-shot Model Attribution Using Invariant Features

The preceding section constructs invariant zones and their corresponding representations $s_m(x)$, which capture model-specific semantic structure while remaining stable under paraphrasing. We now use these representations for model attribution.

**Attribution Formulation.** Let $\mathcal{F} = \{m_1, \ldots, m_K\}$ denote a collection of language models. Given an input prompt $x$, the attribution task is to infer its originating model $m \in \mathcal{F}$ using invariant zone representations. For each reference model $m$, we compute its signature $s_m(x)$, and for a test model we compute $s_t(x)$. Attribution is then performed by similarity matching $\hat{m} = \arg\max_{m \in \mathcal{F}} \mathrm{Sim}\big(s_t(x), s_m(x)\big)$, where $\mathrm{Sim}(\cdot, \cdot)$ denotes a similarity measure such as cosine similarity. This formulation corresponds to a zero-shot prototypical [19] style classification scheme in invariant representation space, as illustrated in Fig. 3. Each model is represented by its invariant zone, and attribution is performed by selecting the closest match.

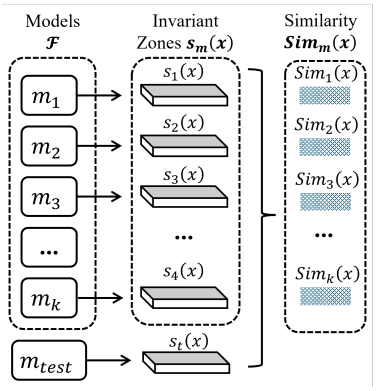


Figure 3: Zero-shot model attribution via invariant zone signatures. Each model $m$ is represented by an invariant signature $s_m(x)$. Given a test model, attribution selects the reference model with highest similarity in invariant space.

**Interpretation as Hypothesis Validation.** Attribution is not posed as a standalone forensic task, but as validation of the proposed geometric framework. If invariant zone representations retain discriminative model identity, this indicates that the discovered invariant structures encode model-specific organization of semantic information. In this way, attribution provides empirical evidence that semantic invariance in language models is a geometric property of latent representations.

## 4 Experiments

We evaluate the proposed geometric framework through a series of analyses that examine the emergence, structure, and functional role of invariant features across models and layers. Core experiments focus on dataset and model setup, invariant zone localization, geometric validation, causal intervention, and attribution, while extended ablations and additional analyses are deferred to the appendix for completeness.

### 4.1 Evaluation Setup

**Dataset Construction.** We construct a dataset of 1,400 semantic groups comprising 57,400 sentences across 13 domains (Table 1), covering diverse topics and linguistic contexts to investigate semantic invariance under controlled SP and SC perturbations. Each group defines a local semantic neighborhood centered on an anchor input. SP variants preserve the underlying meaning while varying lexical

| Component | Count |
|---|---|
| Semantic groups | 1,400 |
| Domains | 13 |
| Anchors/group | 1 |
| SP/group | 20 |
| Soft SC/group | 10 |
| Hard SC/group | 10 |
| Total sentences | 57,400 |

Table 1: Dataset composition. Each of the 1,400 semantic groups contains one anchor, 20 SP, 10 soft SC, and 10 hard SC variants.

form, grammatical structure, and writing style, including formal/casual and question/statement reformulations. Soft SC variants modify an attribute or semantic slot while retaining the original topic, whereas hard SC variants introduce domain- or topic-level semantic changes. For example, given the anchor "What is the official language of France?", an SP variant is "Which language holds official status in France?", a soft SC variant is "What is the national currency of France?", and a hard SC variant is "What is the longest river in South America?". Each SC example includes edit-type and source/target-domain metadata, enabling perturbation-specific analysis. Additional construction details and examples are provided in Appendix E.

To prevent data leakage, each of the 1,400 semantic groups is divided into disjoint discovery and validation subsets, each containing 10 SP and 10 SC variants (5 soft and 5 hard). The discovery subset is used exclusively for invariant subspace identification and layer selection through generalized eigenvalue analysis, while the held-out validation subset is used for downstream evaluation, including tangent-direction validation, representation intervention, projection analysis, and model attribution. This design preserves the local nature of our framework: each group defines an anchor-centered semantic neighborhood with its own invariant subspace. The consistent results across groups support the existence of repeatable local geometric structure rather than a universal invariant projection.

| Model | Family | Params |
|---|---|---|
| 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 |

Table 2: Language models used for evaluation, spanning diverse families, scales, and training paradigms.

**Models Analyzed.** We evaluate nine transformer-based language models listed in Table 2, spanning diverse architectural families, parameter scales (4B–16B), and training paradigms. The suite includes models trained under different objectives and pipelines, including instruction tuning, chat alignment, and mixture-of-experts architectures. This diversity allows us to assess whether invariant structure is a consistent property of representation geometry across model families, rather than an artifact of a specific architecture, scale, or training procedure.

### 4.2 Localization of Invariant Feature Zones

We investigate where invariant structure emerges across model depth by analyzing the layer-wise generalized eigenvalue spectrum defined in Section 3.2. The goal is to identify layers in which semantic-changing variation is maximally separated from semantic-preserving variation, corresponding to the emergence of invariant feature zones. Since invariant zones are locally defined for each semantic group and model pair, the following analysis presents one representative example from Llama, illustrating how such zones emerge across depth within a specific local semantic neighborhood. For each layer $\ell$, we compute the generalized eigenvalues $\{\lambda_i^{(\ell)}\}_{i=1}^{k}$ from Eq. 2, which quantify the ratio between semantic-changing and semantic-preserving variation along each direction. Large eigenvalues indicate directions that are stable under paraphrasing but responsive to semantic change, and therefore correspond to invariant features. We analyze the distribution of the top-$k$ eigenvalues across layers to characterize how invariant structure evolves with depth.

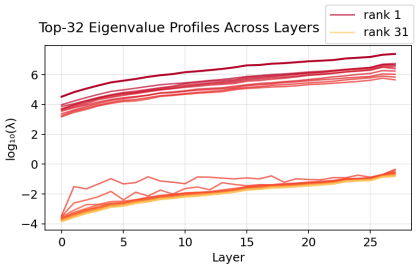


Figure 4: Per-layer eigenvalue profiles for $k = 32$ invariant directions (log scale). Each line denotes an eigenvector rank. Eigenvalues increase with depth, and higher-ranked directions remain dominant, indicating a stable and ordered decomposition of invariant signal across layers.

**Invariant structure becomes more pronounced in mid-to-late layers.** Figure 4 shows the per-layer eigenvalue profiles on a log scale. The leading eigenvalues generally increase with depth, indicating stronger separation between semantic-changing and semantic-preserving variation. However, invariant zones are selected using eigenvalues normalized by the total positive eigenvalue mass within each layer, rather than raw magnitudes alone, reducing bias toward deeper layers. The relative magnitudes of the leading eigenvalues reveal a small set of dominant semantic-sensitive directions, with the strongest separation emerging in mid-to-late layers. This observation is consistent with the local

geometric hypothesis in Section 3.1, suggesting that semantic invariance is concentrated in specific depth regions rather than uniformly distributed across layers. These layers are subsequently used to define invariant zones.

### 4.3 Tangent-Direction Validation of Invariant Geometry

We empirically validate the geometric decomposition introduced in Section 3.2 by testing whether paraphrase-induced variation concentrates in the nuisance subspace, while semantic displacement lies predominantly outside it. This experiment directly evaluates the geometric structure discovered via contrastive sensitivity decomposition.

Let the nuisance component at layer $\ell$ be defined as the Euclidean orthogonal complement of the invariant subspace. Given $U_\ell^{\text{inv}} \in \mathbb{R}^{d\times k}$, the nuisance component of a representation is obtained via the projection operator $(I - U_\ell^{\text{inv}} U_\ell^{\text{inv}\top})$. We do not explicitly construct a nuisance basis, as lower-ranked generalized eigenvectors are not guaranteed to form a stable or Euclidean-orthogonal complement in high-dimensional settings. For each semantic group $g = \{x_1, \ldots, x_{n_g}\}$, we compute latent representations $z_i = \Phi_\ell(x_i)$ and define the group centroid $\bar{z}_g = \frac{1}{n_g}\sum_{i=1}^{n_g} z_i$. For two semantic groups $g$ and $h$, we define the semantic displacement $\Delta_{gh} = \bar{z}_g - \bar{z}_h$. We then evaluate how much of this displacement lies within the nuisance subspace using the projection energy $E_{\text{nuis}} = \left\|(I - U_\ell^{\text{inv}} U_\ell^{\text{inv}\top})\,\Delta_{gh}\right\|_2^2 / \|\Delta_{gh}\|_2^2$. Under the local geometric hypothesis, semantic-preserving variations span the nuisance subspace, while semantic differences lie in complementary directions. In this case, the projection operator $(I - U_\ell^{\text{inv}} U_\ell^{\text{inv}\top})$ removes the semantic component of $\Delta_{gh}$, leaving little residual energy. Consequently, the projected vector $(I - U_\ell^{\text{inv}} U_\ell^{\text{inv}\top})\Delta_{gh}$ has small norm, leading to $E_{\text{nuis}} \approx 0$. Geometrically, $E_{\text{nuis}}$ measures the fraction of semantic displacement energy aligned with nuisance directions. When this value approaches zero, it implies that $\Delta_{gh}$ is nearly orthogonal to the nuisance subspace, and therefore lies primarily in the complementary semantic subspace. This behavior directly reflects the tangent-space decomposition in Eq. 1, where semantic and nuisance variations occupy distinct, approximately orthogonal directions in the local representation geometry. Additional implementation details are provided in Appendix J.

| Model | G1 | G4 | G6 |
|---|---|---|---|
| Mistral | 0.0189 | 0.0213 | 0.0198 |
| Llama3 | 0.0127 | 0.0141 | 0.0135 |
| Qwen2.5 | 0.0158 | 0.0172 | 0.0164 |
| Gemma | 0.0169 | 0.0184 | 0.0176 |
| Phi4-mini | 0.0215 | 0.0231 | 0.0223 |
| GLM4 | 0.0112 | 0.0126 | 0.0119 |
| DeepSeek | 0.0105 | 0.0118 | 0.0112 |
| InternLM2 | 0.0121 | 0.0134 | 0.0128 |
| Falcon3 | 0.0116 | 0.0129 | 0.0123 |

Table 3: Nuisance projection energy $E_{\text{nuis}}$ across models and groups. Low values indicate that semantic displacement lies largely outside the dominant nuisance-sensitive directions.

**Semantic displacement lies predominantly outside the nuisance subspace.** Table 3 reports $E_{\text{nuis}}$ across multiple models and semantic groups. We consistently observe low nuisance projection energy ($\approx 0.01$–$0.02$), indicating that semantic displacement is concentrated in the invariant subspace. This pattern is consistent across models and groups, supporting the proposed local geometric separation between semantic and nuisance directions. Additional comparisons with PCA and cPCA baselines are provided in Appendix D.2.

### 4.4 Ablation Study: Causal Validation via Representation Intervention

We evaluate whether the invariant subspace identified in Section 3.2 causally encodes semantic content through representation-level interventions. Rather than relying solely on correlational analysis, these experiments directly perturb the invariant component of the representation and measure the resulting change in model outputs under both within-group and between-group conditions. Given the layer-wise decomposition $z = z_{\text{inv}} + z_{\text{nuis}}$ obtained from Eq. 3, we construct interventions by replacing the invariant component of an anchor representation while keeping the nuisance component fixed. In practice, interventions are implemented at the token level using the final-token hidden state $h_\ell^{(T)}(x)$, which directly determines next-token prediction in causal language models. We therefore set $z_\ell(x) = h_\ell^{(T)}(x)$, so that $h_\ell^{(T)}(x) = h_{\text{inv}} + h_{\text{nuis}}$. For an anchor hidden state $h^{(a)}$ and donor hidden state $h^{(d)}$, the intervened representation is $h^{\text{int}} = h^{(a)} + \alpha\big(h_{\text{inv}}^{(d)} - h_{\text{inv}}^{(a)}\big)$, which is then propagated through the remaining layers. The intervention effect is measured using the KL divergence between the original and intervened next-token distributions, $\text{KL}(\ell) = D_{\text{KL}}(p_{\text{base}} \,\|\, p_{\text{int}})$ .

We consider three intervention settings. In the within-group SP setting, the donor is a paraphrase from the same semantic group. In the within-group SC setting, the donor is semantically different but remains within the same semantic neighborhood. In the between-group setting, the donor is drawn from a completely different semantic group. Under the geometric hypothesis, SP interventions should produce minimal output change, while SC and between-group interventions should induce substantially larger shifts due to semantic inconsistency.

| **Model** | **Between** KL↑ | **Within-SP** KL↓ | **Within-SC** KL↑ |
|---|---|---|---|
| Mistral-7B | 0.974 | 0.036 | 0.341 |
| Llama-3-8B | 0.986 | 0.040 | 0.360 |
| Gemma-7B | 0.979 | 0.042 | 0.382 |
| Qwen2.5-7B | 0.984 | 0.047 | 0.401 |
| GLM-4-9B | 0.983 | 0.034 | 0.337 |
| DeepSeek-MoE | 0.985 | 0.052 | 0.427 |
| InternLM2-7B | 0.984 | 0.039 | 0.354 |
| Phi-4-mini | 0.981 | 0.045 | 0.391 |
| Falcon3-7B | 0.986 | 0.049 | 0.418 |

Table 4: Invariant-component intervention results. Between-group swaps use donors from a different semantic group; within-group swaps use SP or SC donors from the same group. Low within-SP KL and high within-SC/between-group KL indicate semantic selectivity of the invariant subspace.

**Invariant-component interventions produce substantially larger output changes under semantic-changing than semantic-preserving perturbations.** Table 4 summarizes KL divergence across all three intervention settings. Within-group SP interventions consistently produce near-zero KL divergence across models, indicating that the discovered invariant representation remains stable under paraphrastic variation. In contrast, within-group SC interventions yield substantially larger KL divergence, indicating that the invariant subspace is sensitive to semantic change even within a locally related semantic region. The effect becomes even stronger in the between-group setting, where invariant-component replacement produces consistently large KL divergence across all models. Together, these results provide evidence consistent with a causal role of the discovered invariant subspace in semantically meaningful model behavior. The separation between invariant- and nuisance-component interventions is further supported by narrow 95% bootstrap confidence intervals across semantic groups (Appendix G) and remains evident in multi-hop reasoning tasks (Appendix I).

### 4.5 Robust Model Attribution via Invariant Features

For each model $m \in \mathcal{F}$, attribution is performed by comparing the invariant representation of a test model to reference model signatures using the method defined in Section 3.3. To assess robustness, we consider three variants of each model: the original base pretrained model (B), a fine-tuned version (FT), and a distilled version (D). Fine-tuned models are obtained via instruction tuning on downstream tasks, while distilled models are trained to approximate teacher outputs with reduced capacity. Importantly, prototype invariant zones are computed from the original models and reused without modification for testing the FT and Dist variants. This protocol isolates whether invariant features reflect intrinsic representation geometry rather than training-specific adaptations. Additional implementation details for fine-tuned and distilled model variants, checkpoint construction, layer-selection protocol, and attribution evaluation are provided in Appendix J.6.

| Model | Group1 | | | Group2 | | | Group3 | | |
|---|---|---|---|---|---|---|---|---|---|
| | B | FT | D | B | FT | D | B | FT | D |
| Mistral | 93.2 | 91.4 | 89.8 | 92.7 | 91.0 | 89.4 | 93.5 | 91.8 | 90.2 |
| Llama3 | 94.1 | 92.5 | 90.9 | 93.8 | 92.1 | 90.5 | 94.4 | 92.9 | 91.3 |
| Gemma | 92.8 | 90.9 | 89.2 | 92.3 | 90.5 | 88.8 | 93.0 | 91.2 | 89.6 |
| Qwen2.5 | 93.6 | 91.8 | 90.1 | 93.1 | 91.3 | 89.7 | 93.9 | 92.0 | 90.5 |
| Phi4-mini | 89.6 | 87.8 | 86.1 | 89.2 | 87.4 | 85.8 | 90.1 | 88.2 | 86.5 |
| GLM4 | 95.2 | 93.6 | 91.9 | 94.8 | 93.1 | 91.5 | 95.6 | 94.0 | 92.3 |
| DeepSeek | 95.8 | 94.2 | 92.6 | 95.3 | 93.7 | 92.0 | 96.1 | 94.5 | 92.9 |
| InternLM2 | 94.9 | 93.3 | 91.7 | 94.5 | 92.9 | 91.3 | 95.2 | 93.6 | 91.9 |
| Falcon3 | 95.5 | 93.9 | 92.2 | 95.0 | 93.4 | 91.8 | 95.8 | 94.2 | 92.5 |

Table 5: Attribution accuracy (%) across Base (B), Fine-tuned (FT), and Distilled (D) variants for three representative groups. Performance remains high and stable under model adaptation.

**Invariant representations enable accurate model attribution and remain robust under fine-tuning and distillation.** Table 5 reports attribution accuracy across models and variants. We observe consistently high accuracy for original base models (typically above $92\%$), indicating that invariant zone representations capture strong model-specific signatures. Attribution performance remains high under both fine-tuning and distillation, with only modest degradation (approximately $2$–$4\%$ across models). These results suggest that the discovered invariant features capture model-specific geometric structure that persists under model adaptation, despite changes in task behavior and model parameterization.

**Contrastive invariant decomposition outperforms variance-based and contrastive baselines in model attribution.** To determine whether attribution performance arises from the proposed invariant geometry rather than generic representation structure, we compare our method against PCA and contrastive PCA (cPCA) using matched subspace dimensionality. PCA selects directions of dominant representation variance, while cPCA identifies directions maximizing the contrast $S_{\mathrm{SC}} - \alpha S_{\mathrm{SP}}$. In contrast, our generalized eigenvalue formulation identifies directions with high SC sensitivity relative to SP variation. Table 6 summarizes accuracy averaged across four model families. Our method consistently outperforms both baselines under all adaptation settings, improving over cPCA by 6.13, 6.68, and 7.60 percentage points for base, fine-tuned, and distilled models, respectively. These results indicate that normalizing semantic-changing variation relative to semantic-preserving sensitivity provides more discriminative model signatures than variance maximization or additive covariance contrast alone. Full model-wise comparisons and experimental details are provided in Appendix D.1.

| Variant | PCA | cPCA | Ours |
|---|---|---|---|
| Base | 86.08 | 89.33 | **95.45** |
| Fine-tuned | 83.98 | 87.53 | **94.20** |
| Distilled | 81.25 | 84.65 | **92.25** |

Table 6: Zero-shot attribution accuracy (%) averaged over four models (Llama, Mistral, Qwen2.5, and Gemma) with matched subspace dimensionality. Unlike Table 5, these results are aggregated over the full evaluation set. Full results are in Appendix D.1.

## 5 Limitation and Discussion

Our framework provides a local geometric characterization of semantic invariance rather than a universal decomposition of language representations. Although our evaluation spans 1,400 semantic groups, 57,400 sentences, and nine language models, the decomposition remains dependent on the quality and coverage of the SP and SC perturbations. The dataset is synthetically constructed, and generalization to naturally occurring, multilingual, and long-context inputs remains to be systematically evaluated.

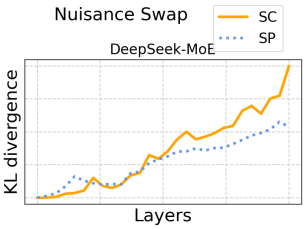


Figure 5: Per-layer KL divergence under nuisance-component swaps for a model. SP and SC curves increase with depth at comparable rates with no consistent separation. The nuisance subspace likely contains a mixture of surface-form and residual semantic information.

While the invariant subspace $z_{\mathrm{inv}}$ exhibits strong semantic selectivity, the complementary nuisance subspace $z_{\mathrm{nuis}}$ does not achieve complete semantic disentanglement. As shown in Fig. 5, swapping nuisance components within the same semantic group produces non-negligible output shifts even for semantic-preserving (SP) donors, indicating that the nuisance representation may retain residual semantic information. Although the two subspaces are orthogonal by construction, the nuisance subspace can still capture a mixture of surface-form variation and residual semantic components. Importantly, this asymmetry does not undermine the primary findings: invariant-component swaps remain stable under SP perturbations while exhibiting strong separation under semantic-changing conditions (Section 4.4), supporting the conclusion that semantic-sensitive information is concentrated in the invariant subspace. Achieving more complete semantic disentanglement remains an important direction for future work.

## 6 Conclusion

We presented a local geometric framework for identifying semantic-invariant features in language model representations through contrastive sensitivity decomposition. Experiments across nine language models demonstrate geometric separation between semantic and nuisance variation, causal sensitivity to semantic changes, and robust zero-shot model attribution. Together, these findings support a local geometric characterization of semantic invariance and its model-specific structure.

## References


[1] K. Ethayarajh, "How contextual are contextualized word representations? comparing the geometry of BERT, ELMo, and GPT-2 embeddings," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 55–65. [Online]. Available: https://aclanthology.org/D19-1006/

[2] I. Tenney, D. Das, and E. Pavlick, "BERT rediscovers the classical NLP pipeline," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 4593–4601. [Online]. Available: https://aclanthology.org/P19-1452/

[3] J. Hewitt and C. D. Manning, "A structural probe for finding syntax in word representations," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4129–4138. [Online]. Available: https://aclanthology.org/N19-1419/

[4] A. Rogers, O. Kovaleva, and A. Rumshisky, "A primer in BERTology: What we know about how BERT works," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 842–866, 2020. [Online]. Available: https://aclanthology.org/2020.tacl-1.54/

[5] E. Reif, A. Yuan, M. Wattenberg, F. B. Viegas, A. Coenen, A. Pearce, and B. Kim, "Visualizing and measuring the geometry of bert," in *Advances in neural information processing systems*, vol. 32, 2019.

[6] R. Bommasani, K. Davis, and C. Cardie, "Interpreting pretrained contextualized representations via reductions to static embeddings," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, Jul. 2020, pp. 4758–4781. [Online]. Available: https://aclanthology.org/2020.acl-main.431/

[7] M. Raghu, J. Gilmer, J. Yosinski, and J. Sohl-Dickstein, "Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[8] A. S. Morcos, M. Raghu, and S. Bengio, "Insights on representational similarity in neural networks with canonical correlation," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[9] S. Kornblith, M. Norouzi, H. Lee, and G. Hinton, "Similarity of neural network representations revisited," in *Proceedings of the 36th International Conference on Machine Learning (ICML)*. PMLR, 2019, pp. 3519–3529.

[10] B. Lester, R. Al-Rfou, and N. Constant, "The power of scale for parameter-efficient prompt tuning," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021. [Online]. Available: https://aclanthology.org/2021.emnlp-main.243/

[11] E. Wallace, S. Feng, N. Kandpal, M. Gardner, and S. Singh, "Universal adversarial triggers for attacking and analyzing NLP," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 2153–2162. [Online]. Available: https://aclanthology.org/D19-1221/

[12] W. Antoun, B. Sagot, and D. Seddah, "From text to source: Results in detecting large language model-generated content," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 2024, pp. 7531–7543.

[13] J. Wang, X. Lu, Z. Zhao, Z. Dai, C.-S. Foo, S.-K. Ng, and B. K. H. Low, "Source attribution for large language model-generated data," 2023. [Online]. Available: https://arxiv.org/abs/2310.00646

[14] A. Dasgupta, A. A. Tanvir, and X. Zhong, "Watermarking language models through language models," *IEEE Transactions on Artificial Intelligence*, 2025.

[15] J. Xu, F. Wang, M. Ma, P. W. Koh, C. Xiao, and M. Chen, "Instructional fingerprinting of large language models," in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. Mexico City, Mexico: Association for Computational Linguistics, Jun. 2024, pp. 3277–3306. [Online]. Available: https://aclanthology.org/2024.naacl-long.180/

[16] C. Fefferman, S. Mitter, and H. Narayanan, "Testing the manifold hypothesis," *Journal of the American Mathematical Society*, vol. 29, no. 4, pp. 983–1049, 2016.

[17] K. Kawaguchi, "Deep learning without poor local minima," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 29, 2016.

[18] A. Choromanska, M. Henaff, M. Mathieu, G. B. Arous, and Y. LeCun, "The loss surfaces of multilayer networks," in *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics (AISTATS)*, ser. Proceedings of Machine Learning Research, vol. 38. PMLR, 2015, pp. 192–204. [Online]. Available: https://proceedings.mlr.press/v38/choromanska15.html

[19] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *Advances in neural information processing systems*, vol. 30, 2017.

[20] R. Taori, I. Gulrajani, T. Zhang, Y. Dubois, X. Li, C. Guestrin, P. Liang, and T. B. Hashimoto, "Stanford alpaca: An instruction-following llama model," https://github.com/tatsu-lab/stanford_alpaca, 2023.

[21] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.

[22] P. Bajaj, D. Campos, N. Craswell, L. Deng, J. Gao, X. Liu, R. Majumder, A. McNamara, B. Mitra, T. Nguyen, M. Rosenberg, X. Song, A. Stoica, S. Tiwary, and T. Wang, "MS MARCO: A human generated machine reading comprehension dataset," in *Proceedings of the Workshop on Cognitive Computation: Integrating Neural and Symbolic Approaches*, 2016.

# Appendix


**A Ablation Study: More Results on Causal Validation via Representation Intervention** **13**

**B Extended Discussion on Limitation: Diagnostic Analysis of Nuisance Subspace** **15**

**C Robustness of Invariant Interventions Across Subspace Dimensions** **17**

**D Comparison with Variance-Based and Contrastive Baselines** **19**

**E Full Dataset Examples** **21**

**F Robustness to Semantic Ambiguity** **22**

**G Statistical Uncertainty on the Expanded Dataset** **22**

**H Effect of Pooling Strategy** **22**

**I Multi-hop Reasoning Validation** **24**

**J Implementation Details** **24**

**K Visualizing the Geometric Separability of Invariant Representations** **28**

**L Broader Impacts** **30**


## A Ablation Study: More Results on Causal Validation via Representation Intervention

We provide additional causal intervention results to further analyze the functional role of the discovered invariant and nuisance subspaces across multiple language models. The experiments include both within-group interventions, where donor representations are drawn from the same local semantic neighborhood, and between-group interventions, where donor representations originate from different semantic groups. Together, these analyses evaluate whether the proposed decomposition selectively controls semantic behavior.

**Experimental Setup.** For each semantic group, invariant directions are identified using the discovery subset by solving the generalized eigenvalue problem in Section 3.2 and selecting the top-$k$ eigenvectors associated with the largest generalized eigenvalues. The top-4 layers are then selected according to the largest leading eigenvalue $\lambda$, corresponding to layers exhibiting the strongest semantic–nuisance separation. All intervention experiments are performed exclusively on the held-out validation subset, with no overlap between discovery and evaluation perturbations.

At each selected layer, token-level hidden states are decomposed into invariant and nuisance components through orthogonal projection onto the discovered subspaces:

$$h = h_{\text{inv}} + h_{\text{nuis}}.$$

Interventions are applied directly to hidden states before propagating the modified activations through the remaining frozen transformer layers.

For within-group interventions, donor representations are sampled from either semantic-preserving (SP) or semantic-changing (SC) perturbations within the same local semantic group. For between-group interventions, donor representations are drawn from a different semantic group, introducing a stronger semantic mismatch between anchor and donor.

| Model | ($A_{\text{nuis}} \leftarrow A_{\text{SP,nuis}}$) | | ($A_{\text{inv}} \leftarrow A_{\text{SP,inv}}$) | | ($A_{\text{nuis}} \leftarrow A_{\text{SC,nuis}}$) | | ($A_{\text{inv}} \leftarrow A_{\text{SC,inv}}$) | |
|---|---|---|---|---|---|---|---|---|
| | KL ↓ | MSE ↓ | KL ↓ | MSE ↓ | KL ↓ | MSE ↓ | KL ↑ | MSE ↑ |
| Mistral-7B-Instruct-v0.3 | 0.124 | 0.581 | 0.036 | 0.401 | 0.109 | 0.612 | 0.341 | 0.951 |
| Meta-Llama-3-8B-Instruct | 0.138 | 0.630 | 0.040 | 0.434 | 0.114 | 0.646 | 0.360 | 0.990 |
| gemma-7b-it | 0.131 | 0.602 | 0.042 | 0.455 | 0.121 | 0.670 | 0.382 | 1.044 |
| Qwen2.5-7B-Instruct | 0.142 | 0.658 | 0.047 | 0.498 | 0.129 | 0.702 | 0.401 | 1.102 |
| glm-4-9b-chat-hf | 0.118 | 0.541 | 0.034 | 0.386 | 0.111 | 0.594 | 0.337 | 0.918 |
| deepseek-moe-16b-chat | 0.149 | 0.691 | 0.052 | 0.523 | 0.138 | 0.741 | 0.427 | 1.184 |
| internlm2-chat-7b | 0.126 | 0.590 | 0.039 | 0.417 | 0.116 | 0.621 | 0.354 | 0.972 |
| Phi-4-mini-instruct | 0.135 | 0.617 | 0.045 | 0.471 | 0.123 | 0.689 | 0.391 | 1.068 |
| Falcon3-7B-Instruct | 0.144 | 0.673 | 0.049 | 0.512 | 0.132 | 0.718 | 0.418 | 1.146 |

Table 7: Within-group causal intervention analysis using invariant and nuisance subspace patching across different language models. Interventions are performed using semantic-preserving (SP) and semantic-changing (SC) donors from the same local semantic group. For nuisance swaps ($A_{\text{nuis}} \leftarrow A_{\text{SP,nuis}}$ and $A_{\text{nuis}} \leftarrow A_{\text{SC,nuis}}$), KL divergence and MSE remain relatively low, indicating limited semantic sensitivity. In contrast, invariant swaps with SC donors ($A_{\text{inv}} \leftarrow A_{\text{SC,inv}}$) produce substantially larger deviations than SP donor swaps ($A_{\text{inv}} \leftarrow A_{\text{SP,inv}}$), demonstrating that the discovered invariant subspace selectively controls semantic behavior. Results are averaged across the top-4 invariant layers selected by the largest generalized eigenvalues.

We independently replace either the invariant or nuisance component while keeping the complementary component fixed. Invariant interventions are defined as

$$h^{\text{int}} = h^{(a)} + \alpha \left( h^{(d)}_{\text{inv}} - h^{(a)}_{\text{inv}} \right),$$

while nuisance interventions are defined as

$$h^{\text{int}} = h^{(a)} + \left( h^{(d)}_{\text{nuis}} - h^{(a)}_{\text{nuis}} \right).$$

Model behavior before and after intervention is evaluated using KL divergence between next-token distributions and logit-space Mean Squared Error (MSE).

| Model | Group 1 | | Group 4 | | Group 6 | |
|---|---|---|---|---|---|---|
| | MSE ↓ | KL ↓ | MSE ↓ | KL ↓ | MSE ↓ | KL ↓ |
| Mistral-7B | 0.021 | 0.084 | 0.024 | 0.091 | 0.022 | 0.087 |
| Llama-3-8B | 0.018 | 0.072 | 0.020 | 0.079 | 0.019 | 0.075 |
| Qwen2.5-7B | 0.019 | 0.078 | 0.003 | 0.033 | 0.020 | 0.080 |
| Gemma-7B | 0.020 | 0.019 | 0.061 | 0.014 | 0.021 | 0.083 |
| Phi-4-mini | 0.024 | 0.031 | 0.026 | 0.042 | 0.025 | 0.097 |
| GLM-4-9B | 0.022 | 0.067 | 0.028 | 0.088 | 0.023 | 0.090 |
| DeepSeek-MoE-16B | 0.025 | 0.081 | 0.030 | 0.095 | 0.027 | 0.102 |
| InternLM2-7B | 0.023 | 0.074 | 0.027 | 0.082 | 0.024 | 0.088 |
| Falcon3-7B | 0.026 | 0.089 | 0.031 | 0.099 | 0.029 | 0.105 |

Table 8: Between-group nuisance-component intervention evaluated using Mean Squared Error (MSE) and KL divergence across models and semantic groups. The invariant component is held fixed while the nuisance component is replaced using a donor from a different semantic group. Low MSE and KL indicate that modifying the nuisance subspace induces only limited changes in the output, suggesting that it does not encode primary semantic content. Results are computed using invariant zones defined by top-4 layers selected via generalized eigenvalue analysis.

**Results.** Tables 7, 8, and 9 summarize the intervention results across all evaluated models.

A consistent asymmetry emerges between invariant and nuisance interventions. For nuisance swaps, both within-group and between-group interventions generally produce relatively small KL divergence and MSE values. Even when nuisance components are replaced using donors from different semantic groups, the resulting output perturbation remains limited. This indicates that the nuisance subspace

does not encode primary semantic identity and has comparatively weak causal influence on semantic behavior.

In contrast, invariant interventions produce substantially larger deviations, particularly when donor representations originate from semantically different inputs. Within-group interventions using SP donors produce relatively small changes, while SC donors induce much larger deviations. This separation becomes even more pronounced under between-group interventions, where replacing the invariant component consistently produces strong shifts in model outputs across all architectures.

Importantly, these trends remain highly consistent across model families, layers, and semantic groups. The results therefore provide complementary causal evidence that the discovered invariant subspace encodes semantically meaningful directions governing model behavior, whereas the nuisance subspace primarily captures semantic-preserving variability and non-selective residual variation.

| Model | Group 1 | | Group 4 | | Group 6 | |
|---|---|---|---|---|---|---|
| | MSE ↑ | KL ↑ | MSE ↑ | KL ↑ | MSE ↑ | KL ↑ |
| Mistral-7B | 0.962 | 0.974 | 0.968 | 0.978 | 0.965 | 0.976 |
| Llama-3-8B | 0.975 | 0.986 | 0.978 | 0.988 | 0.976 | 0.987 |
| Qwen2.5-7B | 0.971 | 0.984 | 0.982 | 0.991 | 0.972 | 0.985 |
| Gemma-7B | 0.966 | 0.979 | 0.980 | 0.986 | 0.967 | 0.980 |
| Phi-4-mini | 0.969 | 0.981 | 0.971 | 0.984 | 0.970 | 0.982 |
| GLM-4-9B | 0.968 | 0.983 | 0.973 | 0.986 | 0.969 | 0.984 |
| DeepSeek-MoE-16B | 0.970 | 0.985 | 0.975 | 0.988 | 0.972 | 0.986 |
| InternLM2-7B | 0.969 | 0.984 | 0.974 | 0.987 | 0.971 | 0.985 |
| Falcon3-7B | 0.972 | 0.986 | 0.976 | 0.989 | 0.974 | 0.987 |

Table 9: Between-group invariant-component intervention evaluated using Mean Squared Error (MSE) and KL divergence across models and semantic groups. The nuisance component is held fixed while the invariant component is replaced using a donor from a different semantic group. High values indicate large and consistent output deviation, suggesting that the invariant subspace encodes semantic identity. Results are computed using invariant zones defined by top-4 layers selected via generalized eigenvalue analysis.

**Discussion.** The contrast between within-group and between-group interventions further clarifies the role of the learned subspaces. Within-group interventions evaluate local semantic stability under paraphrastic variation, while between-group interventions introduce stronger semantic displacement with minimal lexical or structural overlap.

The observed asymmetry between invariant and nuisance interventions suggests that the proposed decomposition captures structured semantic geometry rather than superficial activation statistics. Perturbations in the invariant subspace induce systematic semantic changes, whereas perturbations in the nuisance subspace produce comparatively weak and non-selective effects. Together, these findings support a causal interpretation of invariant zones as semantically meaningful latent directions that persist across models and semantic contexts.

# B Extended Discussion on Limitation: Diagnostic Analysis of Nuisance Subspace

**Token-Level Intervention Protocol.** We follow the same token-level intervention framework described in Appendix A. For each input, we extract the hidden state at the final token position and apply the decomposition $h = h_{\text{inv}} + h_{\text{nuis}}$. Interventions are performed at layer $\ell$ by modifying one component while keeping the other fixed, followed by forward propagation to obtain output logits.

**Nuisance Swap Intervention.** To analyze the behavior of the nuisance subspace, we perform interventions that replace the nuisance component of an anchor representation with that of a donor, while keeping the invariant component unchanged. Donor inputs are drawn from either semantic-preserving (SP) or semantic-changing (SC) sets, and the effect is measured using KL divergence between the original and intervened output distributions.

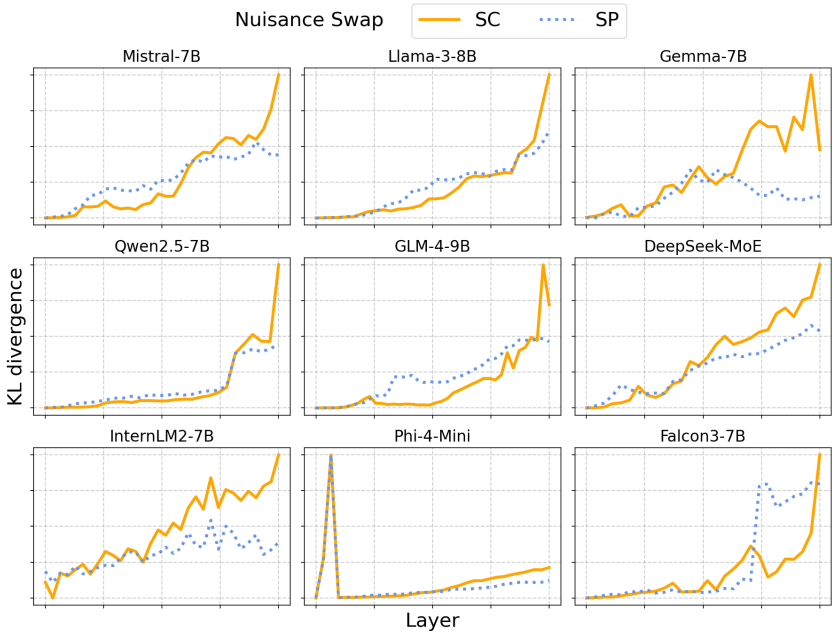


Figure 6: Per-layer KL divergence for nuisance component swaps across all nine models ($k = 32$, $\alpha = 1.0$), arranged as a $3 \times 3$ grid. Orange (solid) curves show SC donor swaps and blue (dotted) curves show SP donor swaps. Both curves increase at comparable rates with no consistent SC-SP separation, indicating that the nuisance subspace does not selectively encode semantic content.

**Observations.** Figure 6 shows the per-layer KL divergence under nuisance swaps across all nine models. In contrast to invariant interventions, SP and SC curves exhibit comparable magnitudes and similar growth patterns across layers, with no consistent separation between the two conditions.

This behavior indicates that perturbations in the nuisance subspace affect model outputs regardless of semantic consistency. Unlike the invariant subspace, where changes align with semantic differences, nuisance perturbations produce output shifts that are not selectively tied to semantic variation.

**Diagnostic Interpretation.** These results provide insight into the limitation observed in Section 5. The decomposition $h = h_{\text{inv}} + h_{\text{nuis}}$ is derived from a data-driven generalized eigenvalue problem that separates directions based on relative sensitivity to semantic-preserving and semantic-changing perturbations. However, geometric orthogonality does not imply complete semantic disentanglement, and the nuisance component may retain residual semantic information.

From a geometric perspective, the tangent-space decomposition $T_z\mathcal{M} = T_z^{\text{semantic}} \oplus T_z^{\text{nuisance}}$ is only approximately realized in learned representations. Finite-sample estimation, nonlinearity of the representation manifold, and imperfect perturbation design introduce mixing between the two subspaces. As a result, the nuisance component may retain weak projections onto semantic directions.

This explains why nuisance swaps produce non-negligible output changes even for SP donors: although dominant semantic directions are removed, residual semantic information persists in the nuisance subspace. Consequently, interventions on $h_{\text{nuis}}$ do not isolate purely non-semantic variation.

**Implications.** This asymmetry highlights a key property of the proposed framework. The invariant subspace provides a selective encoding of semantic information, as demonstrated by its strong causal effect under SC interventions and stability under SP perturbations. In contrast, the nuisance subspace is non-selective: it captures a mixture of surface-form variation and residual semantic components, leading to output changes that are insensitive to semantic alignment.

Importantly, this limitation does not invalidate the main conclusions. Rather, it clarifies that the proposed decomposition should be interpreted as a dominant-direction separation rather than a perfectly disentangled factorization. The invariant subspace concentrates semantic structure, while the nuisance subspace captures complementary but partially entangled variation.

**Summary.** Together with the within-group and between-group intervention results, this analysis supports a consistent causal picture: semantic information is concentrated in invariant directions that govern model behavior, while nuisance directions contribute secondary, non-selective variation. Achieving more complete semantic disentanglement of the nuisance decomposition remains an important direction for future work.

## C Robustness of Invariant Interventions Across Subspace Dimensions

This section evaluates the robustness of the proposed decomposition across models, semantic groups, and subspace dimensions $k \in \{1, 10, 32\}$. We report within-group per-layer KL divergence under invariant and nuisance component interventions, averaged across all concept groups. The intervention protocol follows Section 4.4, where either the invariant or nuisance component of an anchor representation is replaced using a donor drawn from the SP or SC set.

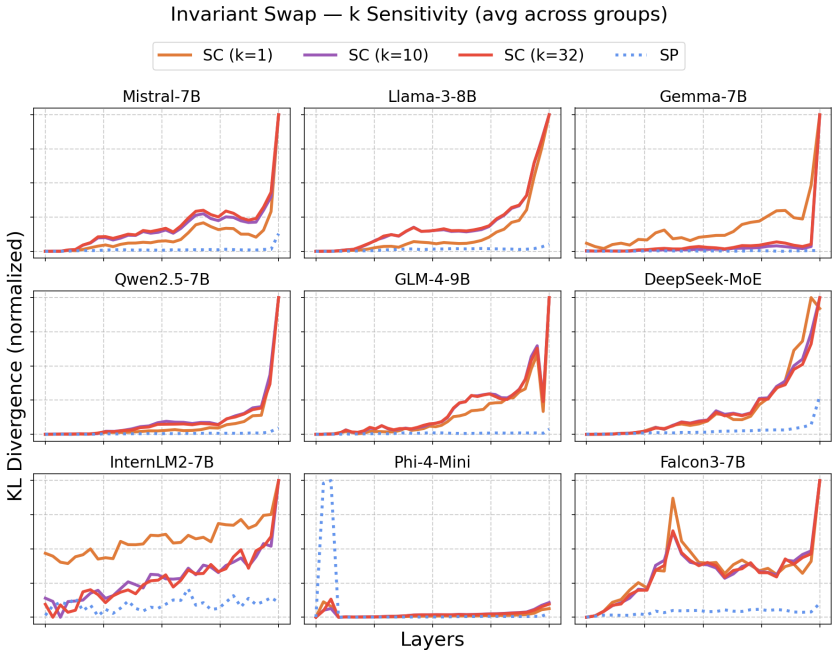


Figure 7: Per-layer KL divergence for invariant component swaps across all nine models, averaged across all 1400 concept groups ($\alpha = 1.0$). Each subplot shows three solid curves corresponding to $k \in \{1, 10, 32\}$ for SC donors, and one dotted curve for SP. SC curves increase consistently with depth while the SP curve remains near zero, suggesting that the invariant subspace selectively encodes semantic content. The three SC curves closely overlap, indicating insensitivity to the choice of $k$.

Figures 7 and 8 present results for invariant and nuisance swaps, respectively. Each figure is arranged as a $3 \times 3$ grid with one subplot per model. Solid curves correspond to SC donor swaps for different values of $k$, and the dotted curve corresponds to SP donor swaps.

**Interpretation.** For invariant swaps (Fig. 7), SC curves increase consistently with depth across all models and all values of $k$, while SP curves remain near zero. This confirms that perturbations in the invariant subspace induce semantically meaningful changes, whereas substitutions with semantically equivalent content leave model outputs largely unchanged. The separation between SC and SP conditions becomes most pronounced in mid-to-late layers, aligning with the invariant zones identified in Section 4.2.

For nuisance swaps (Fig. 8), both SC and SP curves increase at comparable rates with no consistent separation. This behavior corroborates the diagnostic analysis in Appendix B, indicating that the nuisance subspace does not selectively encode semantic content. Instead, it contributes non-selective variation that affects outputs regardless of semantic alignment.

Together, these results reinforce a consistent causal interpretation: invariant components control semantically structured behavior, while nuisance components introduce secondary, non-selective perturbations.

**Insensitivity to Subspace Dimension $k$.** Across both invariant and nuisance interventions, the SC curves for $k \in \{1, 10, 32\}$ closely overlap, indicating that the results are largely insensitive to the choice of subspace dimension. This robustness suggests that the dominant invariant structure is concentrated in a low-dimensional subspace, and that the leading generalized eigenvectors already capture the primary semantic directions.

This observation is consistent with the eigenvalue spectrum in Section 4.2, where a small number of directions exhibit significantly larger eigenvalues. Increasing $k$ primarily introduces weaker components with diminishing influence on downstream behavior.

From a theoretical perspective, the framework requires only that a subspace exists that separates semantic-preserving and semantic-changing variation, rather than a precise specification of its dimensionality. The empirical stability across $k$ therefore indicates that the identified invariant structure reflects an intrinsic geometric property of the representation, rather than a result of hyperparameter tuning.

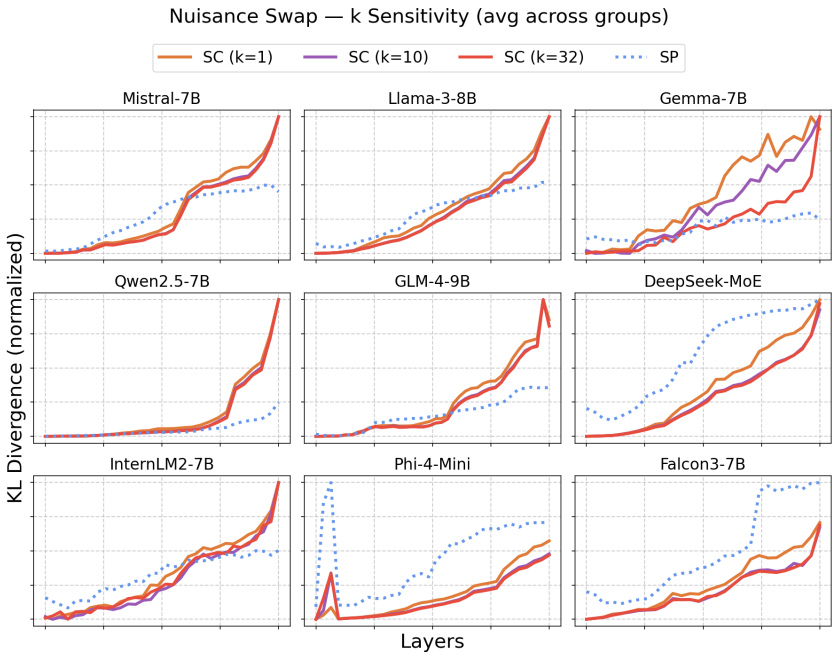


Figure 8: Per-layer KL divergence for nuisance component swaps across all nine models, averaged across all 1400 concept groups ($\alpha = 1.0$). Each subplot shows three solid curves for SC donors at $k \in \{1, 10, 32\}$ and one dotted curve for SP donors. Both SC and SP curves increase at comparable rates with no consistent separation, indicating that the nuisance subspace does not selectively encode semantic content regardless of the choice of $k$.

**Model-Level Variation.** While the overall pattern is consistent across architectures, Phi-4-Mini exhibits a minor deviation under invariant swaps, with slightly elevated SP responses in early-to-mid layers. This may be attributed to its smaller parameter scale, suggesting that semantic abstraction may be less sharply localized across depth. Despite this variation, the SC–SP separation remains evident, and the overall qualitative behavior is consistent with larger models. We report this as a tentative observation for future investigation.

# D Comparison with Variance-Based and Contrastive Baselines

## D.1 Attribution Baseline Comparison

**Experimental Setup.** To evaluate whether invariant zone signatures capture model-specific semantic structure beyond trivial architectural separability, we compare the proposed invariant representation ($Z_{inv}$) against two baselines: a PCA-based latent decomposition and contrastive PCA (cPCA). PCA provides a strong variance-maximizing control that preserves dominant activation energy without explicitly separating semantic-preserving and semantic-changing variation. cPCA, in contrast, explicitly contrasts the two: given the SC covariance $S_{sc}$ and the SP covariance $S_{sp}$, it extracts the leading eigenvectors of $S_{sc} - \alpha S_{sp}$, where $\alpha$ controls the penalty on variation shared with the semantic-preserving background.

For each model family, invariant signatures are constructed from the invariant zones identified through the generalized eigenvalue formulation in Section 3.2, while the PCA and cPCA baselines are computed using the same hidden representations and the same subspace dimensionality. Attribution is performed using held-out validation perturbations only, ensuring that discovery perturbations used for invariant zone identification are never reused during evaluation.

We evaluate three settings: (1) original base models, (2) supervised fine-tuned variants, and (3) knowledge-distilled variants. Fine-tuned and distilled checkpoints are described in Appendix J.6. In all cases, attribution is performed using invariant signatures extracted from the original base model without re-estimating invariant zones for adapted variants. This design tests whether invariant geometric structure remains discriminative under model adaptation.

| Model | Base Model | | | Fine-Tuning | | | Knowledge Distillation | | |
|---|---|---|---|---|---|---|---|---|---|
| | PCA | cPCA | $Z_{inv}$ | PCA | cPCA | $Z_{inv}$ | PCA | cPCA | $Z_{inv}$ |
| Llama-3.1-8B | 87.4 | 89.7 | **96.8** | 85.1 | 87.4 | **95.3** | 82.7 | 84.8 | **93.9** |
| Mistral-7B | 86.2 | 87.5 | **95.7** | 84.6 | 88.2 | **94.8** | 81.9 | 84.7 | **92.6** |
| Qwen2.5-7B | 85.8 | 92.2 | **95.1** | 83.4 | 89.1 | **93.8** | 80.6 | 86.3 | **91.7** |
| Gemma-7B | 84.9 | 87.9 | **94.2** | 82.8 | 85.4 | **92.9** | 79.8 | 82.8 | **90.8** |

Table 10: Zero-shot model attribution accuracy (%) using invariant-zone signatures derived from the proposed contrastive decomposition ($Z_{inv}$), PCA, and contrastive PCA (cPCA) baselines with matched subspace dimensionality. Evaluation is performed across base, fine-tuned, and knowledge-distilled variants using held-out validation perturbations only. The proposed invariant representation consistently outperforms both PCA and cPCA across all settings, suggesting that the identified invariant zones capture model-specific semantic organization beyond dominant activation variance alone.

**Results.** Table 10 summarizes attribution accuracy across model families and adaptation settings. The proposed invariant representation consistently outperforms both PCA and cPCA across all experiments, with improvements remaining stable under both fine-tuning and knowledge distillation. cPCA improves over PCA in every setting but remains below $Z_{inv}$.

Importantly, PCA already captures dominant activation variance and therefore implicitly preserves architectural and scale-related information. The consistent improvement of $Z_{inv}$ over PCA suggests that attribution performance is not solely explained by superficial architectural separability or raw activation statistics. Instead, the contrastive invariant decomposition identifies directions that are simultaneously semantically stable and discriminative across model families. The remaining gap to cPCA, which also contrasts SC and SP variation, suggests that the improvement is not only due to using contrastive information but may also reflect how the generalized eigenvalue formulation normalizes SC variation by SP variation.

These results support the interpretation that invariant zones encode persistent model-specific semantic organization rather than merely reflecting global activation magnitude or dimensionality differences. The fact that attribution remains effective under fine-tuning and distillation further suggests that invariant structure captures stable geometric characteristics that survive substantial model adaptation.

### D.2 Projection Energy Comparison

**Experimental Setup.** We evaluate whether the proposed decomposition separates semantic and nuisance variation more effectively than variance-based representations across multiple language model families, including LLaMA, Gemma, Mistral, Qwen, and DeepSeek. Hidden states are extracted from every transformer layer using the final-token representation.

For each semantic group, semantic-preserving (SP) and semantic-changing (SC) perturbations are used to construct covariance statistics for the generalized eigenvalue decomposition described in Section 3.2. The nuisance subspace is defined using the same process as in section 4.3. To evaluate semantic leakage into nuisance directions, we compute semantic centroids for two different semantic groups and measure the normalized projection energy of their semantic displacement vector onto the nuisance subspace. Lower nuisance energy indicates stronger semantic–nuisance separation.

We compare the proposed method against a PCA baseline computed from the same hidden representations and matched subspace dimensionality. PCA covariance matrices are constructed from the union of anchor, SP, and SC samples. All evaluations use held-out validation perturbations only, ensuring that discovery perturbations used for invariant zone identification are never reused during evaluation.

| Layer | LLaMA | | Gemma | | Mistral | | Qwen | | DeepSeek | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Ours | PCA | Ours | PCA | Ours | PCA | Ours | PCA | Ours | PCA |
| 0 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| 1 | 0.0025 | 0.7320 | 0.0055 | 0.6440 | 0.0028 | 0.5702 | 0.0054 | 0.5998 | 0.0031 | 0.5940 |
| 2 | 0.0100 | 0.6208 | 0.0022 | 0.5532 | 0.0069 | 0.5607 | 0.0139 | 0.5746 | 0.0016 | 0.5160 |
| 3 | 0.0038 | 0.6892 | 0.0047 | 0.5359 | 0.0024 | 0.6096 | 0.0048 | 0.5158 | 0.0030 | 0.5661 |
| 4 | 0.0050 | 0.5584 | 0.0054 | 0.3913 | 0.0035 | 0.4085 | 0.0048 | 0.4519 | 0.0021 | 0.5061 |
| 5 | 0.0109 | 0.4182 | 0.0031 | 0.3576 | 0.0037 | 0.4069 | 0.0023 | 0.5534 | 0.0034 | 0.4001 |
| 6 | 0.0056 | 0.5097 | 0.0039 | 0.3178 | 0.0039 | 0.5340 | 0.0045 | 0.5970 | 0.0036 | 0.3861 |
| 7 | 0.0042 | 0.5357 | 0.0026 | 0.5032 | 0.0017 | 0.5070 | 0.0025 | 0.4837 | 0.0014 | 0.3528 |
| 8 | 0.0080 | 0.5216 | 0.0040 | 0.4180 | 0.0050 | 0.4622 | 0.0073 | 0.2989 | 0.0034 | 0.3053 |
| 9 | 0.0088 | 0.5324 | 0.0038 | 0.3932 | 0.0065 | 0.4638 | 0.0065 | 0.2803 | 0.0056 | 0.2591 |
| 10 | 0.0082 | 0.5330 | 0.0058 | 0.3944 | 0.0075 | 0.4595 | 0.0026 | 0.3821 | 0.0037 | 0.2693 |
| 11 | 0.0126 | 0.5927 | 0.0054 | 0.3827 | 0.0052 | 0.4674 | 0.0099 | 0.3565 | 0.0019 | 0.3138 |
| 12 | 0.0063 | 0.5857 | 0.0052 | 0.3879 | 0.0038 | 0.4737 | 0.0089 | 0.3121 | 0.0032 | 0.3366 |
| 13 | 0.0097 | 0.5625 | 0.0086 | 0.3677 | 0.0050 | 0.4850 | 0.0041 | 0.3216 | 0.0061 | 0.3069 |
| 14 | 0.0106 | 0.5745 | 0.0043 | 0.3772 | 0.0047 | 0.5006 | 0.0073 | 0.3188 | 0.0052 | 0.3320 |
| 15 | 0.0085 | 0.5698 | 0.0063 | 0.3646 | 0.0028 | 0.5248 | 0.0026 | 0.3224 | 0.0050 | 0.2934 |
| 16 | – | – | 0.0055 | 0.3707 | 0.0057 | 0.5418 | 0.0063 | 0.3205 | 0.0040 | 0.3201 |
| 17 | – | – | 0.0050 | 0.3902 | 0.0029 | 0.5669 | 0.0069 | 0.3453 | 0.0020 | 0.3524 |
| 18 | – | – | 0.0058 | 0.3943 | 0.0012 | 0.5632 | 0.0028 | 0.3630 | 0.0067 | 0.3434 |
| 19 | – | – | 0.0060 | 0.4044 | 0.0032 | 0.5646 | 0.0051 | 0.4171 | 0.0067 | 0.3419 |
| 20 | – | – | 0.0046 | 0.3911 | 0.0033 | 0.5314 | 0.0065 | 0.4379 | 0.0052 | 0.3554 |
| 21 | – | – | 0.0030 | 0.4161 | 0.0025 | 0.5444 | 0.0050 | 0.4555 | 0.0035 | 0.3679 |
| 22 | – | – | 0.0036 | 0.4396 | 0.0025 | 0.5413 | 0.0066 | 0.4401 | 0.0045 | 0.3634 |
| 23 | – | – | 0.0051 | 0.4168 | 0.0039 | 0.5410 | 0.0036 | 0.4523 | 0.0025 | 0.3919 |
| 24 | – | – | 0.0029 | 0.4507 | 0.0045 | 0.5403 | 0.0038 | 0.4400 | 0.0047 | 0.3754 |
| 25 | – | – | 0.0053 | 0.4403 | 0.0046 | 0.5471 | 0.0034 | 0.4550 | 0.0042 | 0.4409 |
| 26 | – | – | 0.0020 | 0.4168 | 0.0027 | 0.5526 | 0.0032 | 0.4457 | 0.0025 | 0.4843 |
| 27 | – | – | 0.0051 | 0.4301 | 0.0037 | 0.5485 | 0.0022 | 0.3956 | 0.0017 | 0.4940 |
| 28 | – | – | – | – | 0.0027 | 0.5334 | – | – | 0.0039 | 0.5009 |
| 29 | – | – | – | – | 0.0018 | 0.5279 | – | – | 0.0022 | 0.5127 |
| 30 | – | – | – | – | 0.0052 | 0.5190 | – | – | 0.0028 | 0.5245 |
| 31 | – | – | – | – | 0.0027 | 0.5206 | – | – | 0.0026 | 0.5273 |

Table 11: Layer-wise comparison of normalized nuisance projection energy ($E_{\text{nuis}}$) between the proposed invariant decomposition and a PCA baseline across multiple language model families. Lower values indicate better semantic–nuisance separation. Evaluation uses held-out validation perturbations only for representative semantic groups ($\text{group}_g = 0$, $\text{group}_h = 1$). Missing layers are denoted by (—). A cPCA comparison is given in Table 12.

**Results.** Table 11 reports full layer-wise nuisance projection energy for representative semantic groups ($\text{group}_g = 0$, $\text{group}_h = 1$). Across all evaluated architectures and layers, the proposed

decomposition consistently produces extremely small nuisance projection energy, typically on the order of $10^{-3}$, while PCA yields substantially larger values ranging from approximately $0.3$ to $0.7$.

These results indicate that the proposed SP/SC-aware decomposition isolates semantic structure substantially more effectively than variance-based PCA decomposition. Importantly, PCA already preserves dominant activation variance and global representation statistics, yet still exhibits large semantic leakage into nuisance directions. The strong and consistent gap therefore suggests that semantic separation is not explained solely by unsupervised variance structure.

The proposed method also remains highly stable across shallow and deep transformer layers, whereas PCA exhibits consistently large nuisance projections throughout the network hierarchy. This behavior further supports the interpretation that semantic structure is more effectively characterized through semantic-preserving and semantic-changing perturbation analysis than through variance maximization alone.

**Comparison with cPCA.** Table 12 provides a representative comparison of nuisance projection energy with cPCA, using the same representations, subspace dimensionality, and evaluation protocol. cPCA substantially improves over PCA, indicating that contrastive information is important. Our method achieves the lowest nuisance energy at all evaluated layers. Nuisance energies on the semantic ambiguity stress test are reported in Table 15 (Appendix F).

| Layer | Ours ↓ | cPCA ↓ | PCA ↓ |
|---|---|---|---|
| 1 | **0.0025** | 0.0323 | 0.7320 |
| 4 | **0.0050** | 0.0676 | 0.5584 |
| 8 | **0.0085** | 0.0691 | 0.8851 |
| 12 | **0.0053** | 0.0590 | 0.9572 |
| 16 | **0.0047** | 0.0152 | 0.9689 |

Table 12: Layerwise nuisance projection energy for LLaMA on the expanded dataset for Ours, cPCA, and PCA. Lower values are better; the best result in each row is shown in bold.

# E Full Dataset Examples

**Dataset Construction.** The dataset consists of 1,400 semantic groups spanning 13 domains, comprising 57,400 sentences in total. Each group is centered on an anchor query and contains 20 semantic-preserving (SP) variants, 10 soft semantic-changing (SC) variants, and 10 hard SC variants. SP variants introduce lexical substitutions, syntactic restructuring, or changes in writing style while preserving the propositional content of the anchor. Soft SC variants modify a semantic attribute or slot while largely retaining the original topic and sentence structure. Hard SC variants introduce a topic- or domain-level semantic change, producing a more substantial departure from the anchor's meaning.

This construction provides controlled local neighborhoods containing both semantic-preserving and semantic-changing directions. The distinction between soft and hard SC variants enables analysis of semantic changes at different levels, from localized modifications to broader topic shifts. SC examples are additionally annotated with edit-type and source/target-domain metadata to support perturbation-specific analyses.

**Discovery and Held-Out Validation Split.** For each anchor, the 20 SP variants and 20 SC variants are divided equally into two disjoint subsets: a discovery subset and a held-out validation subset. Each subset contains 10 SP variants, 5 soft SC variants, and 5 hard SC variants. The discovery subset is used exclusively to estimate the local covariance operators, identify invariant subspaces through generalized eigenvalue analysis, and select representation layers. The held-out subset is reserved for downstream evaluations, including projection analysis, causal intervention, and model attribution.

The split is performed within each anchor neighborhood, ensuring that subspace identification and evaluation use disjoint perturbation instances while remaining centered on the same local semantic region. This design evaluates whether the identified invariant directions generalize beyond the perturbation instances used for their estimation, without assuming transfer across unrelated semantic neighborhoods.

**Representative Examples.** Table 13 presents representative examples from ten thematic groups. For each anchor, the first row contains SP and SC variants from the discovery subset, while the second row contains variants from the held-out validation subset. The displayed SC examples illustrate localized semantic changes; the full dataset additionally includes hard SC variants sampled from different topics or domains.

**Perturbation Coverage and Evaluation Scope.** The dataset is designed to characterize local representation geometry under controlled semantic variation rather than to exhaustively sample natural language. Within each anchor neighborhood, the SP and SC perturbations provide contrasting directions for estimating and evaluating semantic invariance. For example, an anchor about domestic cats may be paired with a soft SC variant concerning their sleeping behavior or a hard SC variant such as *"Describe the migration patterns of monarch butterflies."*

Because semantic preservation is not always binary, we further examine clear SP cases (e.g., harmless typos and grammatical reformulations), ambiguous cases involving scope, punctuation, or syntactic attachment, and clear SC cases involving negation, role reversal, or quantifier changes. The corresponding results are reported in Appendix F.

The consistency of geometric measurements and intervention results across nine language models and diverse semantic groups provides evidence that the observed patterns extend beyond individual anchor queries. Statistical reliability is further assessed using 95% bootstrap confidence intervals over the 1,400 semantic groups, as reported in Appendix G (Table 16). Nevertheless, the controlled construction does not fully capture the diversity of naturally occurring language variation, particularly multilingual inputs and long-context dependencies, which remain directions for future evaluation.

## F Robustness to Semantic Ambiguity

The boundary between SP and SC perturbations can depend on context: changes in surface form may preserve meaning, alter meaning, or introduce ambiguity. We construct an additional stress test with three categories: clear SP perturbations involving harmless typos or grammatical reformulations; ambiguous perturbations involving scope, punctuation, or attachment; and clear SC perturbations involving negation, role reversal, or quantifier changes. Table 14 illustrates these categories. We estimate the decomposition from clear SP and SC variation and evaluate its behavior on the stress-test inputs, including ambiguous cases. Using nuisance projection energy, our method achieves lower values than PCA and cPCA in every reported setting (Table 15). These results extend the baseline comparison to inputs where surface form and meaning are closely coupled. The decomposition captures dominant semantic-sensitive directions rather than imposing a rigid binary boundary on every perturbation.

## G Statistical Uncertainty on the Expanded Dataset

We estimate uncertainty in the mean intervention effects across 1,400 semantic groups for each model. At each layer, we perform 1,000 bootstrap trials, each sampling 1,400 groups with replacement and recomputing the mean nuisance- and invariant-component intervention metrics. We obtain 95% confidence intervals from the 2.5th and 97.5th percentiles of the bootstrap distribution. Table 16 reports the across-group means and confidence intervals on the expanded dataset at representative layers (3, 5, and 8). For every model and every reported layer, invariant-component interventions produce larger MSE and KL changes than nuisance-component interventions, with confidence intervals narrow relative to the separation between the two intervention types.

## H Effect of Pooling Strategy

We compare mean pooling and last-token pooling across four models at layer 3 using the same decomposition and intervention protocol. As shown in Table 17, both strategies produce larger SC than SP intervention effects in KL divergence and MSE, preserving semantic selectivity. Last-token pooling consistently yields lower SP disturbance and nuisance projection energy, although its absolute SC effects are also smaller. We therefore retain last-token pooling as a conservative choice that limits sensitivity to meaning-preserving changes while remaining responsive to semantic changes.

| Group | Anchor Query | SP Variant | SC Variant |
|---|---|---|---|
| Cats | Describe the hunting behavior of domestic cats. | Describe the hunting habits of domestic cats.<br>Detail how house cats pursue and capture prey. | Describe the sleeping behavior of domestic cats.<br>Describe the nocturnal behavior of domestic cats. |
| France | What is the official language of France? | Which language is officially recognized in France?<br>Identify the language that serves as the official one in France. | What is the capital city of France?<br>What is the official flower of France? |
| Python | Why did Python become the dominant language for data science? | What factors led to Python becoming the dominant language in data science?<br>Explain the reasons behind Python's leading position in data science. | Why did Python become popular in academia?<br>Why did Python become popular for scientific computing? |
| Photosynthesis | What are the pros and cons of engineering crops for enhanced photosynthesis? | What are the advantages and disadvantages of engineering crops for improved photosynthesis?<br>Discuss the benefits and costs of engineering crops to improve photosynthetic performance. | What are the pros and cons of engineering crops for drought resistance?<br>What are the pros and cons of engineering crops for improved carbon sequestration? |
| Democracy | Describe the role of free speech in a democratic society. | What role does free speech play in a democratic society?<br>Explain the function of free expression in democratic systems. | Describe the role of free press in a democratic society.<br>Describe the role of voting rights in a democratic society. |
| Einstein | What is Einstein's most famous theory? | Which theory is Albert Einstein best known for?<br>Identify the theory for which Albert Einstein is most widely recognized. | What is Einstein's nationality?<br>What is Einstein's most famous thought experiment? |
| Climate Change | Describe the effect of climate change on global sea levels. | What impact does climate change have on global sea levels?<br>Explain how climate change affects sea levels on a global scale. | Describe the effect of climate change on Arctic ecosystems.<br>Describe the effect of climate change on global monsoon patterns. |
| Calculus | How do you find the derivative of a composite function? | How is the derivative of a composite function calculated?<br>Explain how to differentiate a function composed of other functions. | How do you evaluate the limit of a function?<br>How do you compute the second derivative of a function? |
| Meditation | Describe the cognitive benefits of regular meditation practice. | What cognitive benefits are associated with regular meditation practice?<br>Explain the mental effects of consistent meditation practice. | Describe the physical benefits of regular meditation practice.<br>Describe the psychological benefits of regular meditation practice. |
| Bread | What is the primary leavening agent in sourdough bread? | What is the main leavening agent in sourdough bread?<br>Identify the ingredient responsible for leavening in sourdough bread. | What is the main ingredient in sourdough bread?<br>What is the typical crust texture of sourdough bread? |

Table 13: Representative SP and SC variants from 1400 thematic groups. For each anchor, the first row shows discovery-subset examples and the second row shows held-out validation examples. Each group contains 20 SP variants and 20 SC variants (10 soft and 10 hard), divided equally between the two subsets. The table illustrates soft SC changes; hard SC variants introduce larger topic- or domain-level shifts.

| Category | Original | Perturbation | Semantic Effect |
|---|---|---|---|
| Clear SP | Alice is taller than Bob. | Bob is shorter than Alice. | Preserves the comparison. |
| Ambiguity | I saw the man who was holding a telescope. | I saw the man with a telescope. | Introduces ambiguity about who used or held the telescope. |
| Clear SC | Not every student submitted the assignment. | No student submitted the assignment. | Changes partial non-submission to universal non-submission. |

Table 14: Illustrative perturbation pairs for the semantic ambiguity stress test.

| Setting | Layer | Ours ↓ | cPCA ↓ | PCA ↓ |
|---|---|---|---|---|
| Ambiguity | 3 | **0.0230** | 0.0718 | 0.1170 |
| Ambiguity | 4 | **0.0084** | 0.0718 | 0.0921 |
| Clear SP | 3 | **0.0122** | 0.0389 | 0.2673 |
| Clear SC | 4 | **0.0123** | 0.1284 | 0.2843 |

Table 15: Nuisance projection energy on the semantic ambiguity stress test, averaged across all evaluated models. Lower values are better; the best result in each row is shown in bold.

# I Multi-hop Reasoning Validation

Multi-hop reasoning requires combining multiple facts or intermediate inferences to reach a conclusion, rather than relying on a single explicit fact. We additionally evaluate 200 multi-hop reasoning examples to test whether the invariant decomposition remains selective when semantic differences require reasoning beyond direct lexical or attribute substitutions. Table 18 shows example instances from this dataset. Using the same decomposition and intervention protocol, we compare the effects of replacing semantic and nuisance components. As shown in Table 19, semantic interventions produce larger MSE and KL changes than nuisance interventions at all reported layers. These results support the conclusion that the discovered invariant directions remain semantically selective in reasoning-intensive settings.

# J Implementation Details

## J.1 Generalized Eigenvalue Regularization and Whitening Equivalence

Since the number of perturbations is much smaller than the representation dimension, the empirical semantic-preserving covariance $S_{\mathrm{SP}}$ is typically rank-deficient. We therefore introduce ridge regularization $B = S_{\mathrm{SP}} + \epsilon I$, with $\epsilon > 0$ ($\epsilon$ is set to a small fraction, e.g., $10^{-4}$), and solve

$$S_{\mathrm{SC}} v = \lambda B v. \tag{4}$$

Because $B$ is symmetric positive definite, its inverse square root $B^{-1/2}$ exists. Substituting $u = B^{1/2} v$ yields the equivalent standard eigenvalue problem

$$B^{-1/2} S_{\mathrm{SC}} B^{-1/2} u = \lambda u, \qquad v = B^{-1/2} u. \tag{5}$$

Thus, our generalized eigenvalue formulation is equivalent to whitening the representation space with respect to regularized SP variation before identifying the dominant SC-sensitive directions. The regularization ensures numerical stability when $S_{\mathrm{SP}}$ is ill-conditioned. The resulting generalized eigenvectors are subsequently Euclidean-orthonormalized, as described in Section 3.2, to construct the invariant projection operator.

## J.2 Cross-Layer Relationship of Invariant Subspaces

Invariant subspaces are estimated independently at each transformer layer, without assuming a fixed basis across depth. Since adjacent layers are connected through nonlinear transformations, their local invariant directions may be approximately related through the layer Jacobian, i.e., $U^{\mathrm{inv}}_{\ell+1} \approx$

| Model | Layer | Nuisance KL | | Invariant KL | | Nuisance MSE | | Invariant MSE | |
|---|---|---|---|---|---|---|---|---|---|
| Mistral-7B | 3 | 0.000294 | ± | 0.183353 | ± | 0.001498 | ± | 0.554108 | ± |
| | | 0.000022 | | 0.000924 | | 0.000840 | | 0.000398 | |
| | 5 | 0.001740 | ± | 0.048431 | ± | 0.001239 | ± | 0.124954 | ± |
| | | 0.000084 | | 0.000880 | | 0.000699 | | 0.000296 | |
| | 8 | 0.000535 | ± | 0.115219 | ± | 0.002914 | ± | 0.538303 | ± |
| | | 0.000022 | | 0.000627 | | 0.000477 | | 0.000858 | |
| Llama-3-8B | 3 | 0.001294 | ± | 0.010279 | ± | 0.003071 | ± | 0.024246 | ± |
| | | 0.000098 | | 0.000647 | | 0.000194 | | 0.001549 | |
| | 5 | 0.006545 | ± | 0.039726 | ± | 0.016768 | ± | 0.091234 | ± |
| | | 0.000497 | | 0.001782 | | 0.001773 | | 0.003889 | |
| | 8 | 0.009954 | ± | 0.109119 | ± | 0.027611 | ± | 0.211765 | ± |
| | | 0.000670 | | 0.003465 | | 0.002654 | | 0.007343 | |
| Qwen2.5-7B | 3 | 0.001692 | ± | 0.141399 | ± | 0.002138 | ± | 0.280972 | ± |
| | | 0.000049 | | 0.000169 | | 0.000237 | | 0.000876 | |
| | 5 | 0.000498 | ± | 0.103953 | ± | 0.001561 | ± | 0.449803 | ± |
| | | 0.000021 | | 0.000996 | | 0.000520 | | 0.000882 | |
| | 8 | 0.002991 | ± | 0.134659 | ± | 0.001339 | ± | 0.024437 | ± |
| | | 0.000074 | | 0.001000 | | 0.000355 | | 0.000509 | |
| Gemma-7B | 3 | 0.000982 | ± | 0.083551 | ± | 0.001444 | ± | 0.528555 | ± |
| | | 0.000041 | | 0.000415 | | 0.000264 | | 0.000336 | |
| | 5 | 0.000252 | ± | 0.174981 | ± | 0.002773 | ± | 0.362715 | ± |
| | | 0.000035 | | 0.000732 | | 0.000624 | | 0.000635 | |
| | 8 | 0.001787 | ± | 0.175533 | ± | 0.002566 | ± | 0.143605 | ± |
| | | 0.000030 | | 0.000438 | | 0.000295 | | 0.000834 | |
| Phi-4-mini | 3 | 0.001135 | ± | 0.150466 | ± | 0.002899 | ± | 0.216349 | ± |
| | | 0.000073 | | 0.000103 | | 0.000279 | | 0.000892 | |
| | 5 | 0.001609 | ± | 0.157141 | ± | 0.002398 | ± | 0.200817 | ± |
| | | 0.000055 | | 0.000147 | | 0.000918 | | 0.000420 | |
| | 8 | 0.000629 | ± | 0.137118 | ± | 0.002453 | ± | 0.483504 | ± |
| | | 0.000069 | | 0.000304 | | 0.000747 | | 0.000699 | |
| GLM-4-9B | 3 | 0.001060 | ± | 0.025708 | ± | 0.002141 | ± | 0.017117 | ± |
| | | 0.000083 | | 0.000593 | | 0.000409 | | 0.000323 | |
| | 5 | 0.002712 | ± | 0.077275 | ± | 0.002567 | ± | 0.074551 | ± |
| | | 0.000082 | | 0.000622 | | 0.000122 | | 0.000403 | |
| | 8 | 0.001142 | ± | 0.181325 | ± | 0.001591 | ± | 0.142025 | ± |
| | | 0.000044 | | 0.000979 | | 0.000384 | | 0.000125 | |
| DeepSeek-MoE-16B | 3 | 0.002609 | ± | 0.147375 | ± | 0.002082 | ± | 0.265895 | ± |
| | | 0.000038 | | 0.000650 | | 0.000829 | | 0.000227 | |
| | 5 | 0.002764 | ± | 0.190571 | ± | 0.002600 | ± | 0.029917 | ± |
| | | 0.000089 | | 0.000611 | | 0.000257 | | 0.000208 | |
| | 8 | 0.001402 | ± | 0.027540 | ± | 0.001993 | ± | 0.117297 | ± |
| | | 0.000064 | | 0.000402 | | 0.000376 | | 0.000493 | |
| InternLM2-7B | 3 | 0.002665 | ± | 0.163942 | ± | 0.001603 | ± | 0.231792 | ± |
| | | 0.000025 | | 0.000931 | | 0.000531 | | 0.000880 | |
| | 5 | 0.000171 | ± | 0.106152 | ± | 0.002800 | ± | 0.209048 | ± |
| | | 0.000056 | | 0.000652 | | 0.000651 | | 0.000898 | |
| | 8 | 0.002509 | ± | 0.037368 | ± | 0.001217 | ± | 0.292352 | ± |
| | | 0.000086 | | 0.000844 | | 0.000796 | | 0.000495 | |
| Falcon3-7B | 3 | 0.000382 | ± | 0.087830 | ± | 0.002887 | ± | 0.117867 | ± |
| | | 0.000039 | | 0.000734 | | 0.000966 | | 0.000723 | |
| | 5 | 0.001994 | ± | 0.102695 | ± | 0.002094 | ± | 0.392787 | ± |
| | | 0.000037 | | 0.000221 | | 0.000239 | | 0.000200 | |
| | 8 | 0.001578 | ± | 0.027555 | ± | 0.002944 | ± | 0.310904 | ± |
| | | 0.000037 | | 0.000999 | | 0.000565 | | 0.000624 | |

Table 16: Mean intervention metrics on the expanded dataset across 1,400 semantic groups at layers 3, 5, and 8, with 95% bootstrap confidence intervals. Results are reported as mean ± confidence interval half-width.

| Model | Pooling | SP KL ↓ | SC KL ↑ | SP MSE ↓ | SC MSE ↑ | Nuisance Energy ↓ |
|---|---|---|---|---|---|---|
| Llama-3-8B | Mean | 0.001190 | 0.065358 | 0.003095 | 0.123798 | 0.061882 |
| | Last token | 0.000037 | 0.000876 | 0.000087 | 0.001567 | 0.013685 |
| Mistral-7B | Mean | 0.002501 | 0.027186 | 0.000094 | 0.004563 | 0.061063 |
| | Last token | 0.000027 | 0.001793 | 0.000089 | 0.000729 | 0.009884 |
| Qwen2.5-7B | Mean | 0.000692 | 0.038182 | 0.000187 | 0.007310 | 0.012205 |
| | Last token | 0.000087 | 0.000445 | 0.000094 | 0.000774 | 0.006332 |
| Gemma-7B | Mean | 0.001159 | 0.008129 | 0.003581 | 0.007889 | 0.066067 |
| | Last token | 0.000108 | 0.000355 | 0.000199 | 0.000385 | 0.044512 |

Table 17: Comparison of mean pooling and last-token pooling at layer 3 under the same decomposition and intervention protocol.

| Type | Supporting Facts | Answer |
|---|---|---|
| **Example 1:** In which country does Maya live? | | |
| Anchor | Maya lives in Brookfield. Brookfield is in Canada. | Canada |
| SP | Maya resides in Brookfield. Brookfield is located in Canada. | Canada |
| SC | Maya lives in Brookfield. Brookfield is in France. | France |
| **Example 2:** Which piece of furniture contains the key? | | |
| Anchor | The key is inside a box. The box is inside a cupboard. | Cupboard |
| SP | A box holds the key. The box is stored inside a cupboard. | Cupboard |
| SC | The key is inside a box. The box is inside a desk. | Desk |
| **Example 3:** Did Aisha finish before Carlos? | | |
| Anchor | Aisha finished before Ben. Ben finished before Carlos. | Yes |
| SP | Ben finished after Aisha but before Carlos. | Yes |
| SC | Aisha finished after Ben. Carlos finished before Ben. | No |

Table 18: Illustrative multi-hop examples with nuisance (SP) and semantic (SC) changes. SP variants preserve the reasoning and answer, whereas SC variants change a fact or relationship and alter the answer. The question remains fixed within each example.

| Layer | Nuisance MSE | Semantic MSE | Nuisance KL | Semantic KL |
|---|---|---|---|---|
| 3 | 0.143 | 0.970 | 0.064 | 0.353 |
| 4 | 0.130 | 0.908 | 0.069 | 0.360 |
| 5 | 0.107 | 0.832 | 0.080 | 0.391 |

Table 19: Intervention results on the additional multi-hop reasoning dataset, averaged across all evaluated models. Semantic-component interventions produce larger changes than nuisance-component interventions at each reported layer.

$\mathrm{orth}(J_\ell U_\ell^{\mathrm{inv}})$. However, this relationship is input-dependent and is not imposed or evaluated in the present framework. Characterizing cross-layer subspace transport remains an interesting direction for future work.

### J.3 Representation Extraction

For each sentence, we extract the hidden state at the final token position from every transformer layer using forward hooks. In causal language models, the final token attends to all preceding tokens and aggregates contextual information from the full input, making it the most semantically rich position for next-token prediction. This yields one $d$-dimensional representation per sentence per layer, where $d$ is the model's hidden dimension (4096 for most models evaluated). All representations are extracted in `float32` precision.

### J.4 Implementation Details for Nuisance Energy Evaluation.

For nuisance energy evaluation, we compute semantic displacement between all ordered semantic group pairs $(g, h)$ with $g \neq h$. Given $N$ semantic groups, this produces $N(N-1)$ ordered semantic displacement pairs. In our experiments, we use a predefined set of semantic groups to generate semantic displacement pairs for each model and layer.

Nuisance energy is evaluated independently at every transformer layer, including the embedding layer and all subsequent transformer blocks. Unless otherwise specified, the invariant subspace dimensionality is fixed to $k = 32$, and the nuisance subspace is defined as its Euclidean orthogonal complement. Hidden representations are extracted from the final-token embedding at each layer and converted to floating-point representations prior to covariance estimation. The sensitivity matrices are computed from anchor-relative displacement vectors $\Delta z = z - z_{\text{orig}}$ as defined in Section 3.2. Semantic displacement vectors are not additionally normalized before projection. Instead, normalization is incorporated implicitly through the energy-ratio formulation, where projected nuisance energy is divided by the total displacement energy.

Table 3 reports some representative groups. Across all evaluated models, the proposed method consistently yields very small nuisance projection energy with low variance across semantic groups and layers, indicating stable separation between semantic and nuisance components throughout the transformer hierarchy.

In contrast, variance-based decomposition methods produce substantially larger nuisance projection energy and greater variability, suggesting weaker disentanglement between semantic and nuisance directions. The consistently low nuisance energy observed in our method therefore reflects stable concentration of semantic variation within the invariant subspace rather than numerical artifacts or isolated layer behavior.

### J.5 Compute Resources

All experiments were conducted on a single NVIDIA RTX 6000 Ada Generation GPU (48 GB VRAM) from a workstation equipped with two such GPUs, an AMD Ryzen Threadripper PRO 7985WX (64 cores), and 252 GB system RAM. A single GPU was used throughout to ensure compatibility with single-GPU reproduction environments. Representation extraction for one model across all layers for one concept group takes approximately 1 minute. The causal intervention experiment takes approximately 3 minutes per model per group. Fine-tuning each base model on the Alpaca dataset for one epoch using LoRA takes approximately 20–40 minutes per model. The total compute for all reported experiments, including eigenvalue analysis, intervention experiments, attribution evaluation, and t-SNE visualization across all models, concept groups, and values of $k$, is estimated at approximately 150 GPU-hours, inclusive of preliminary and repeated runs during development.

### J.6 Fine-Tuned and Distilled Model Variants

To evaluate the robustness of invariant representations under model adaptation, we construct fine-tuned and distilled variants of all nine base models, yielding 18 additional checkpoints (9 fine-tuned, 9 distilled) for a total of 27 model variants evaluated in the attribution experiments.

**Checkpoint Sources.** All adapted model checkpoints are trained by the authors from the original base model weights obtained from HuggingFace Hub. Fine-tuned variants are produced by applying supervised fine-tuning with Low-Rank Adaptation (LoRA) on the Alpaca instruction-following dataset [20], which contains approximately 52,000 instruction-response pairs. Fine-tuning uses cross-entropy loss on the response tokens, with the pretrained model weights frozen and only the LoRA adapter parameters updated. Distilled variants are produced by training a smaller student model to approximate the output distribution of the corresponding teacher on the same Alpaca dataset, using response-level knowledge distillation where the student minimizes the KL divergence between its logits and those of the teacher. Student architectures are selected based on proximity to the teacher family where possible: Gemma-2B for Mistral and Gemma, TinyLlama-1.1B-Chat for Llama, Qwen-1.5B for Qwen, and Phi-3-Mini for the remaining five models. Table 20 summarizes the distillation pairings and parameter scales.

| Teacher Model | Student Model | Teacher Params | Student Params |
|---|---|---|---|
| Mistral-7B-Instruct-v0.3 | Gemma-2B | 7B | 2B |
| Gemma-7B-IT | Gemma-2B | 7B | 2B |
| Llama-3-8B-Instruct | TinyLlama-1.1B-Chat | 8B | 1.1B |
| Qwen2.5-7B-Instruct | Qwen-1.5B | 7B | 1.5B |
| GLM-4-9B-Chat | Phi-3-Mini | 9B | 3.8B |
| DeepSeek-MoE-16B-Chat | Phi-3-Mini | 16B | 3.8B |
| InternLM2-7B-Chat | Phi-3-Mini | 7B | 3.8B |
| Phi-4-Mini-Instruct | Phi-3-Mini | 4B | 3.8B |
| Falcon3-7B-Instruct | Phi-3-Mini | 7B | 3.8B |

Table 20: Distilled model variants used for attribution robustness evaluation. Each student is trained on the Alpaca dataset to approximate the output distribution of the corresponding teacher via KL divergence minimization. Fine-tuned variants (not shown) are produced for all nine base models using the same Alpaca dataset with standard supervised fine-tuning.

**Signature Construction and Layer Selection.** Invariant zone signatures are computed exclusively from the original base model for each model family. The same signatures and layer selections are reused without modification when evaluating fine-tuned and distilled variants. This protocol is intentional: it tests whether the geometric structure discovered in the base model persists under adaptation, rather than re-discovering potentially adapted structure in each variant separately. We acknowledge that this means layer selection is performed on base model representations and may not reflect the optimal layers for adapted variants. However, since attribution is performed by matching a test variant against a base model prototype, the relevant question is whether the base model's invariant zones remain discriminative after adaptation — which is precisely what the results in Table 5 evaluate. Layer selection does not use model identity as a signal and is performed independently per model family prior to any attribution comparison, so no identity leakage occurs through this step.

All 27 model variants are evaluated using invariant zone signatures derived solely from the original base model checkpoints. This design isolates whether invariant representations encode persistent geometric structure that survives model adaptation, rather than being specific to a particular training state.

## K Visualizing the Geometric Separability of Invariant Representations

We visualize that the invariant subspace identified by our framework yields geometrically separable representations across models. We use t-SNE [21] to visualize representation geometry in two settings: our structured custom dataset and an external corpus of queries sampled from MS MARCO [22], a large-scale information retrieval benchmark consisting of real web search queries. MS MARCO queries are drawn from a fundamentally different distribution than our custom SP/SC dataset—they are natural, unstructured, and span a broad range of topics without controlled semantic variation—making them a suitable testbed for evaluating whether the invariant subspace generalizes beyond the data used to derive it.

**Custom Dataset.** Figure 9 shows a t-SNE visualization of invariant subspace projections for all nine models on a single concept group ($k = 32$). For each model, the best layer is selected independently as the layer where the top-$k$ generalized eigenvectors account for the largest fraction of invariant signal. To ensure comparability across layers, eigenvalues at each layer are normalized by the sum of positive eigenvalues at that layer. This normalization emphasizes the relative concentration of invariant signal within each layer rather than absolute magnitude, which increases with depth and would otherwise bias selection toward later layers. Hidden states at the selected layer are collected for all sentences in the group (1 anchor, 10 SP, 10 SC), and each representation is projected into the $k$-dimensional invariant coordinate space via $c = U_{\text{inv}}^{\top} h$. The resulting coordinates are aggregated across models and reduced to 2D using t-SNE for visualization. Despite the limited number of samples per model, the representations form visually separable clusters by model identity, suggesting that the invariant subspace captures model-specific geometric structure.

**MS MARCO Queries.** To evaluate generalization beyond the structured SP/SC dataset, we sample 200 queries from the MS MARCO training split and extract hidden state representations for each

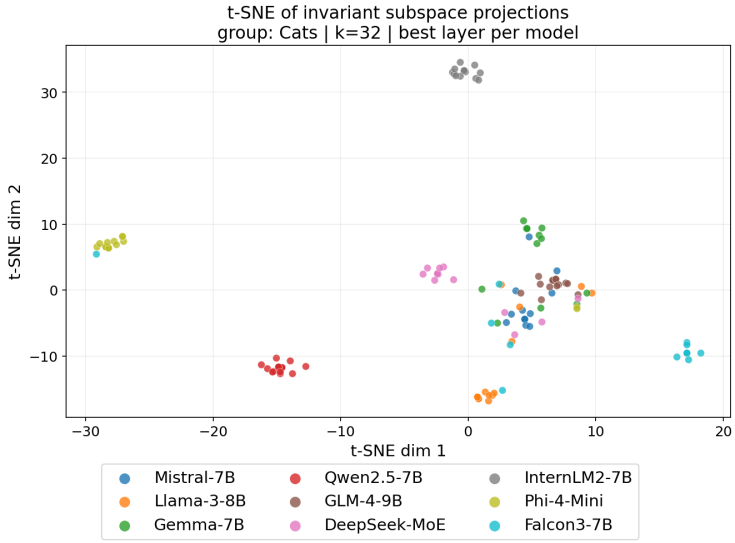


Figure 9: t-SNE visualization of invariant subspace projections ($k = 32$) for all nine models on a single concept group. Each point represents one sentence, colored by model identity.

model at its best layer. The best layer is selected as the one with the highest mean normalized eigenvalue fraction averaged across all concept groups. This cross-group averaging ensures that the selected layer reflects consistently strong invariant structure across diverse semantic domains rather than being tuned to a single concept. Critically, $U_{\text{inv}}$ is derived from the custom SP/SC dataset and held fixed; it is applied to MS MARCO representations without recomputation, testing whether invariant subspaces discovered on controlled data transfer to natural language queries from a different distribution.

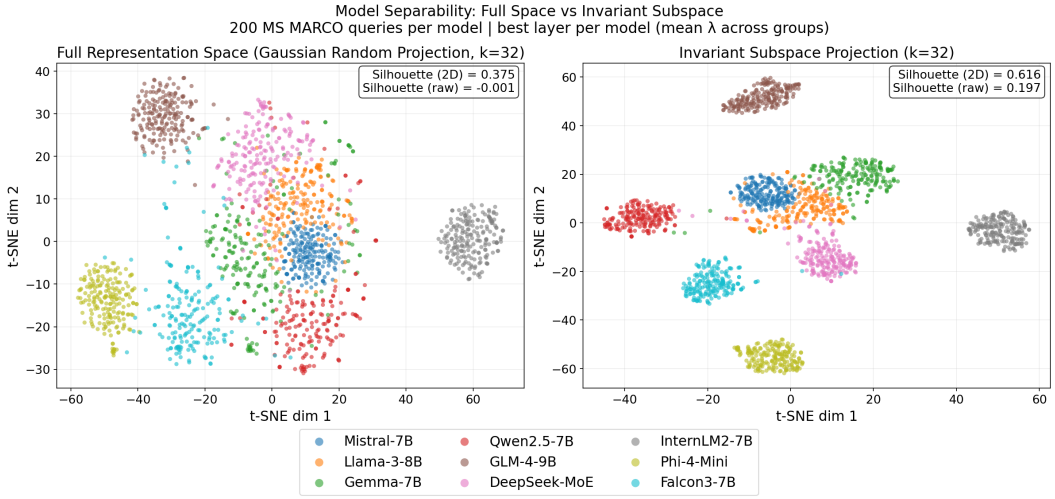


Figure 10: Side-by-side t-SNE comparison of model separability using 200 MS MARCO queries per model ($k = 32$). *Left:* full representation space reduced via Gaussian random projection. *Right:* invariant subspace projection. Silhouette scores are shown for both the 2D t-SNE output and the raw $k$-dimensional coordinates.

Figure 10 compares two conditions. The left panel reduces each model's raw $d$-dimensional hidden states to $k = 32$ dimensions via Gaussian random projection, a distance-preserving linear transformation that treats all directions uniformly and introduces no semantic bias, before applying t-SNE. The right panel applies t-SNE directly to invariant subspace coordinates $c = U_{\text{inv}}^{\top} h \in \mathbb{R}^{32}$, which already share a common dimensionality across models. To quantify separability beyond visual inspection, silhouette scores are computed both on the 2D t-SNE embeddings and on the original $k$-dimensional representations, with both values reported in the figure. The invariant projection exhibits consis-

tently higher separability, indicating that the contrastive subspace concentrates model-discriminative structure while suppressing nuisance variation. This is consistent with the geometric formulation in Section 3.2, where invariant directions are selected to maximize semantic-changing variation relative to semantic-preserving variation.

## L Broader Impacts

This work studies the internal representation geometry of large language models and proposes a framework for identifying invariant latent features that encode semantic information. As a primarily foundational contribution, the goal is to improve scientific understanding of how language models organize meaning and maintain consistency under paraphrastic variation.

A potential positive impact of this work lies in model transparency and interpretability. By providing a principled method to analyze invariant structure, the framework may support better diagnostic tools for understanding model behavior, detecting failure modes, and evaluating robustness under distribution shift. In addition, the use of invariant representations for model attribution may contribute to provenance tracking, intellectual property protection, and forensic analysis of generated content, which are increasingly important in the deployment of generative models.

At the same time, the ability to identify model-specific signatures from internal representations may introduce potential risks. Attribution techniques could be used to infer proprietary information about models or to reverse-engineer aspects of their internal structure without authorization. Furthermore, improved understanding of invariant features may enable adversaries to design attacks that specifically target or manipulate these representations, potentially affecting model reliability or enabling evasion of detection systems.

More broadly, as with many advances in representation learning, the proposed methods are dual-use. While they can enhance robustness, transparency, and accountability, they may also be misused in adversarial settings. We do not release model-specific fingerprints or tools intended for unauthorized identification, and we emphasize that responsible use of such techniques should follow appropriate legal and ethical guidelines.

Overall, we believe the primary impact of this work is to advance the scientific understanding of representation geometry in language models. We encourage future work to further study safeguards, robustness, and responsible deployment of techniques that analyze or utilize invariant latent structure.

## NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Please refer to the abstract and the last paragraph of Section 1 for the enumeration of the contributions. The claims are supported by the experiments in Section 4 and the appendix.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations are discussed in Section 5, which addresses the partial entanglement of the nuisance subspace, the controlled, synthetically constructed evaluation dataset, and the local nature of the framework. The appendix further discusses model-level variation and robustness to hyperparameter choices.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: The paper does not include formal theorems or proofs. Instead, the proposed geometric framework is formulated as a local hypothesis and operationalized empirically. The mathematical formulations and their underlying assumptions are explicitly stated in Section 3.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: All models used are publicly available on HuggingFace Hub. The dataset construction procedure, covariance estimation, generalized eigenvalue formulation, intervention protocol, and hyperparameter settings are described in Section 4 and Appendix J. Additional implementation details, including numerical stabilization and representation extraction, are provided in Appendix J.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Core code and evaluation data are provided in the supplemental material, together with instructions sufficient to reproduce the main experiments. The dataset construction procedure and evaluation protocol are described in Section 4 and Appendix E. All base models are publicly available via HuggingFace Hub under their respective licenses.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Dataset construction, model suite, representation extraction, covariance estimation, subspace dimension $k$, intervention hyperparameters ($\alpha$, swap formulation), and layer selection procedure are all specified in Section 4 and Appendix J. Fine-tuning and distillation details are provided in Appendix J.6.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

Justification: We report 95% bootstrap confidence intervals for intervention metrics across 1,400 semantic groups, using 1,000 bootstrap resamples. The evaluation procedure and results are described in Appendix G.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Compute resources are described in Appendix J.5.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research analyzes internal representations of publicly available language models and introduces no new data collection, human subjects, or deployment pipeline. The work conforms with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.

- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Broader impacts are discussed in Appendix L, covering both positive impacts (model transparency, interpretability, provenance tracking) and potential negative impacts (reverse engineering of proprietary model structure, adversarial exploitation of invariant representations), along with a discussion of responsible use.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release new pre-trained models, scraped datasets, or image generators. All base models used are already publicly available. The evaluation dataset consists of author-constructed prompts with no sensitive content.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All nine base models are cited with their original publications and are accessed via HuggingFace Hub under their respective publicly available licenses. The Alpaca dataset is cited [20] and MS MARCO is cited [22]. Model-specific license terms are available on their respective HuggingFace model cards.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The evaluation dataset introduced in this paper is documented in Section 4.1 and Appendix E, including the construction procedure, semantic group definitions, SP/SC variant generation, and discovery/validation split protocol. A representative anonymized subset of the dataset is included in the supplemental material.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs are the subject of analysis in this paper, not a component of the proposed methodology. LLM assistance was limited to writing support, editing, formatting, and minor brainstorming, and did not contribute to the core scientific framework, mathematical formulation, or experimental results.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.