# Is Inference Mediated by Distinct Semantic Structures in LLMs? A Mechanistic Interpretation

**Anonymous ACL submission**


## Abstract

Predicting a label correctly does not necessarily require representing the operation that produces it. Transformer representations are known to carry label-level information, but whether they encode semantic operations producing those labels is unclear. We investigate this in Natural Language Inference using controlled premise-hypothesis pairs that differ by a single semantic transformation. Using layer-wise activations, we estimate operation-level subspaces via SVD and test their causal relevance through activation steering in four open-weight decoder models. Transformation effects are decodable with 84.8-99% accuracy and occupy partially distinct but overlapping subspaces, exceeding random-subspace baselines. Steering experiments show that these directions causally influence predictions, though steerability varies across models; cross-operation steering further reveals structured interference and a dissociation between subspace selectivity and cross-operation independence. These findings indicate that the models encode not only *that* a hypothesis relates to a premise but also, in part, *how* it does so, implying that mechanistic analysis and control should operate at the level of semantic operations rather than predicted labels alone. [1]


## 1 Introduction

Understanding how transformer models represent semantic relationships requires going beyond what a model predicts to how its internal computation is organised (Aljaafari et al., 2026; Olah et al., 2020). Since natural language inference (NLI) labels can result from different underlying semantic operations (Bowman et al., 2015; Williams et al., 2018), it provides a controlled setting for asking whether models represent only the final relation label or also the transformation that produces it (see App. A for a formal linguistic and logical grounding of this distinction); a question with direct consequences for mechanistic interpretability, evaluation, and control (Geiger et al., 2021).

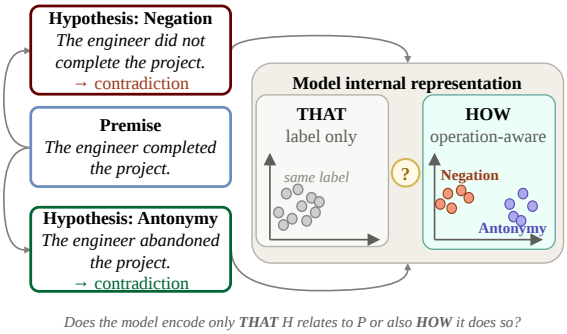


Figure 1: **Transformation-level ambiguity in entailment representations.** A single premise can be paired with hypotheses produced by different semantic operations. We ask whether the model represents only the shared label or also the transformation that produces it.

Fig. 1 illustrates this ambiguity: the premise *The engineer completed the project*, both negation *The engineer did not complete the project*, and antonymy *The engineer abandoned the project* produce a contradiction label despite involving distinct semantic operations. Since the training objective does not directly require these operations to be distinguished, the shared label may obscure whether their internal representations are transformation-specific.

Existing work motivates this question from several directions, but does not address it directly. Probing studies show that semantic and relational properties are often linearly decodable from hidden states (Tenney et al., 2019; Hewitt and Manning, 2019; Belinkov, 2022), but decodability alone does not establish that the model uses this information (Hewitt and Liang, 2019; Voita and Titov, 2020). Interventional work on NLI has identified causally relevant natural-logic features (Geiger et al., 2020;

[1] Code and supplementary materials are available at [anonymisedforreview].

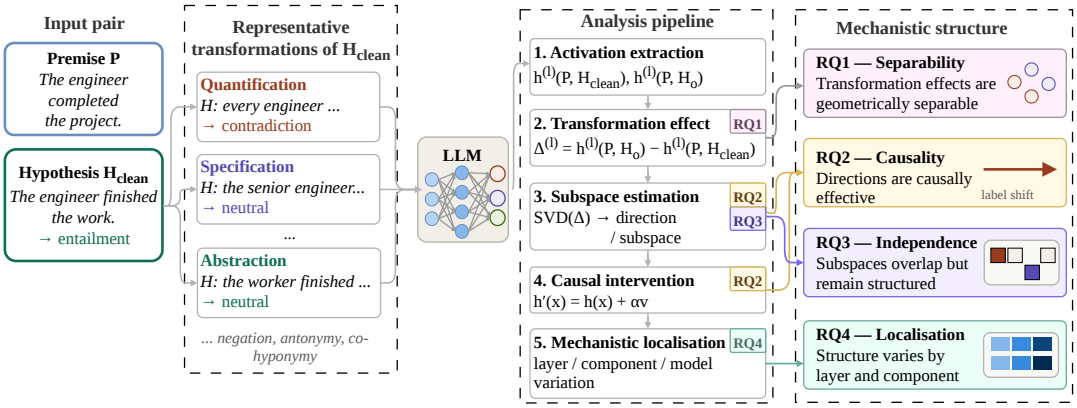


Figure 2: **Method overview.** Controlled transformations are applied to the hypothesis in a premise-hypothesis pair. We extract activation differences $\Delta^{(l)} = h^{(l)}(P, H_o) - h^{(l)}(P, H_{\text{clean}})$ at each layer, estimate operation-specific subspaces $\mathcal{S}_o^{(l)}$ using SVD, and test derived directions through bidirectional activation steering. RQ labels indicate which analysis each stage supports.

Rozanova et al., 2023, 2024), while representation-geometry and activation-steering studies show that semantic directions can be both geometrically consistent and causally influential (Park et al., 2023; Hernandez et al., 2024; Marks and Tegmark, 2024; Turner et al., 2023; Zou et al., 2023). However, these lines of work do not establish whether a *class* of meaning-altering operations shares low-dimensional geometric structure, or whether that structure causally contributes to NLI predictions.

To investigate this, we construct a controlled 1,800-pair contrastive NLI dataset in which each clean premise-hypothesis input is paired with a semantic variant differing by a single meaning-altering operation, spanning six operations in total (Fig. 2). This pairing isolates transformation-induced activation differences, which we analyse through four questions: **(RQ1)** ***Separability:*** whether activation differences from distinct transformation types occupy geometrically distinguishable regions; **(RQ2)** ***Causality:*** whether directions derived from these differences systematically shift model predictions under targeted interventions; **(RQ3)** ***Independence:*** how subspace selectivity relates to cross-operation interference, and whether these properties are dissociable; and **(RQ4)** ***Localisation:*** at which layers and components transformation-dependent structure emerges and label-level evidence is amplified. We study six transformations, Negation, Antonymy, Quantification, Abstraction, Specification, and Co-hyponymy, across four open-weight decoder models at the 1.7-4B scale.

The results provide mechanistic evidence that models encode not only label outcomes but also, in part, the operations that produce them. *Transformation effects are linearly decodable with* 84.8-99% *accuracy across models*, well above random subspace baselines, and *activation steering confirms their causal relevance in three of four models*. Operations differ in how selectively their effects are concentrated within their own subspaces, and *this selectivity is dissociable from cross-operation independence*: an operation with moderate subspace selectivity can produce stronger cross-operation interference than the operation with the highest selectivity. These findings indicate that *label-level analyses undercharacterise model behaviour*, and that *the relevant structure is defined at the level of semantic operations rather than outputs alone*. We summarise our contributions as follows:

- A mechanistic framework combining contrastive transformation-pair construction, operation-specific subspace identification, bidirectional activation steering, and logit attribution to analyse semantic transformations as geometric effects in model representations.
- Empirical evidence that transformation activation differences are *linearly decodable* with high accuracy and *causally steerable*, showing that models encode not only label-level outcomes but also, in part, the operations that lead to them.
- Evidence of a dissociation between subspace selectivity and cross-operation independence, and an asymmetry between forward and inverse steering, both constraining mechanistic accounts of transformation-level representation.

## 2 Related Work

**Probing and representation analysis.** Probing studies show that syntactic, semantic, and relational properties are decodable from transformer representations (Tenney et al., 2019; Jawahar et al., 2019; Hewitt and Manning, 2019), but a successful probe shows that information is *present*, not that the model *uses* it; control-task and information-theoretic refinements sharpen the methodology without closing this gap (Belinkov, 2022; Hewitt and Liang, 2019; Voita and Titov, 2020). Our work adds causal evidence: transformation-level structure is not only decodable but behaviourally active under targeted steering.

**Mechanistic interpretability.** The circuits framework (Olah et al., 2020; Elhage et al., 2021) decomposes transformer computation into interpretable subgraphs; subsequent work has identified components associated with indirect object identification (Wang et al., 2023), modular arithmetic (Nanda et al., 2023), and syntactic agreement (Finlayson et al., 2021), with automated discovery (Conmy et al., 2023) and causal abstraction (Geiger et al., 2021; Wu et al., 2023) scaling these analyses. Rather than identifying circuits that produce a label, we characterise whether *transformation-level effects* form structured geometric objects in activation space that can be isolated and causally manipulated.

**Activation steering and representation engineering.** Contrastive activation addition (Turner et al., 2023; Rimsky et al., 2024), inference-time intervention (Li et al., 2023), representation engineering (Zou et al., 2023), and contrast-consistent search (Burns et al., 2023) show that semantic properties correspond to steerable directions in representation space; Zhao et al. (2026) further demonstrate that SVD-identified subspaces can be ablated to disentangle surface features from reasoning, a close methodological precedent of our approach. We extend this paradigm from broad behavioural properties to six fine-grained semantic operations, enabling direct comparison of transformation-specific directions and their asymmetric induction and reversal behaviour.

**NLI and semantic phenomena.** NLI has served as a central evaluation task for language understanding (Bowman et al., 2015; Williams et al., 2018), with natural logic providing a formal account of the underlying entailment relations (MacCartney and Manning, 2009) and analyses revealing model sensitivity to annotation artefacts (Gururangan et al., 2018), syntactic heuristics (McCoy et al., 2019), and specific phenomena such as negation and lexical entailment (Geiger et al., 2020; Nie et al., 2020). These works often operate at the level of observable outputs; our contribution is orthogonal, investigating the internal representational geometry of the semantic operations that produce NLI predictions.

**Representation geometry.** Linear structure in embeddings (Mikolov et al., 2013) has motivated a line of work showing that entity relations (Hernandez et al., 2024), in-context learning tasks (Todd et al., 2024), and high-level semantic properties (Park et al., 2023; Marks and Tegmark, 2024) occupy consistent directions in transformer representation space. We extend this line from individual relations or binary properties to *operation-level* transformations between sentence pairs, providing causal evidence that the identified geometric structure is behaviourally relevant.

## 3 Methodology

### 3.1 Setup and Notation

Let $\mathcal{X}$ denote the space of NLI inputs, where each $x = (p, h) \in \mathcal{X}$ is a premise-hypothesis pair, and let $f_\theta$ be a pretrained transformer with $L$ layers. At each layer $l$, we extract final-token activations from three components: the residual stream $\mathbf{h}^{(l)}_{\texttt{resid}}(x) \in \mathbb{R}^d$, the attention output $\mathbf{h}^{(l)}_{\texttt{attn}}(x)$, and the MLP output $\mathbf{h}^{(l)}_{\texttt{mlp}}(x)$, following the standard transformer decomposition (Elhage et al., 2021). We study six semantic operations $o \in \mathcal{O}$: Negation, Antonymy, Quantification, Abstraction, Specification, and Co-hyponymy; definitions and examples are given in Table 1 and their formal grounding in compositional semantics and natural logic is provided in App. A. Each operation defines a transformation function $T_o : \mathcal{X} \to \mathcal{X}$ that modifies the hypothesis of an input $x_i$, yielding a contrastive pair $(x_i, T_o(x_i))$. We assume that $T_o$ changes only the intended semantic dimension of the hypothesis; this paired-isolation assumption is enforced through controlled data generation (§4), with operation selection details in App. B.1.

### 3.2 Transformation Effects and Subspace Estimation

Let the activation delta induced by applying operation $o$ to input $x_i$ at layer $l$ (Fig. 2, **RQ1-RQ3**)

Table 1: **Semantic operations studied,** with linguistic characterisation and representative example pairs. For each operation, a shared premise $P$ is fixed and the hypothesis $H$ is modified: $H_{\text{clean}}$ is entailed by $P$, and $H_o$ denotes the hypothesis after applying operation $o$. The resulting label changes from ENTAILMENT to CONTRADICTION (Negation, Antonymy, Quantification) or NEUTRAL (Abstraction, Specification, Co-hyponymy).

| **Operation** | **Characterisation** | **Example** |
|---|---|---|
| *Shared premise: P = The company hired three new engineers last month.* | | |
| *Clean hypothesis:* $H_{\text{clean}}$ *= The company hired new engineers.* | | (ENTAILMENT) |
| Negation | Logical reversal via negation marker | $H_o$: *The company did not hire new engineers.* |
| Antonymy | Lexical substitution with antonym | $H_o$: *The company fired new engineers.* |
| Abstraction | Shift from specific to general term | $H_o$: *The company hired new staff.* |
| Specification | Shift from general to specific term | $H_o$: *The company hired new software engineers.* |
| Quantification | Modification of quantifier scope or strength | $H_o$: *The company hired no new engineers.* |
| Co-hyponymy | Substitution with co-hyponym (same hypernym) | $H_o$: *The company hired new designers.* |

be:

$$\Delta_o^{(l)}(x_i) = \mathbf{h}^{(l)}(T_o(x_i)) - \mathbf{h}^{(l)}(x_i) \tag{1}$$

Collecting $N$ deltas yields $\mathbf{D}_o^{(l)} \in \mathbb{R}^{N\times d}$. Following Zhao et al. (2026), operation-specific structure is identified via truncated singular value decomposition of the mean-centred matrix $\tilde{\mathbf{D}}_o^{(l)} = \mathbf{D}_o^{(l)} - \bar{\boldsymbol{\delta}}_o^{(l)}$, where $\bar{\boldsymbol{\delta}}_o^{(l)}$ is the column-wise mean of $\mathbf{D}_o^{(l)}$:

$$\tilde{\mathbf{D}}_o^{(l)} = \mathbf{U}\,\boldsymbol{\Sigma}\,\mathbf{V}^\top \tag{2}$$

The top-$k$ right singular vectors define an orthonormal basis $\mathbf{B}_o^{(l)} \in \mathbb{R}^{k\times d}$ (row-stacked) for the operation subspace $\mathcal{S}_o^{(l)} \subset \mathbb{R}^d$. For any $\boldsymbol{\delta} \in \mathbb{R}^d$, the orthogonal projection onto $\mathcal{S}_o^{(l)}$ is:

$$\text{proj}_{\mathcal{S}_o^{(l)}}(\boldsymbol{\delta}) = \mathbf{B}_o^{(l)\top}\mathbf{B}_o^{(l)}\,\boldsymbol{\delta} \tag{3}$$

This subspace view is a local approximation: applying $T_o$ is modelled through the activation difference it induces at layer $l$, without assuming that semantic transformations act as globally linear maps. The approximation is motivated by evidence that some relational and truth-conditional properties are linearly recoverable in transformer activations (Park et al., 2023; Hernandez et al., 2024; Marks and Tegmark, 2024), and by activation-steering work showing that contrastive activation differences can yield behaviourally effective directions (Turner et al., 2023; Rimsky et al., 2024). We therefore use SVD to estimate the dominant shared directions of $\Delta_o^{(l)}$, interpreting $\mathcal{S}_o^{(l)}$ as the layer-local geometric structure through which operation $o$ is expressed; assumptions and failure modes are discussed in App. C.4.

### 3.3 Representational Structure

The extent to which transformation effects induce geometrically distinct structure in representation space (**RQ1**) is evaluated using two complementary measures.

**Selectivity.** For each subspace $\mathcal{S}_o^{(l)}$, activation deltas from all operations are projected into $\mathcal{S}_o^{(l)}$ and the selectivity ratio is computed:

$$\rho_o^{(l)} = \frac{\mathbb{E}_{x\sim o}\left[\left\|\text{proj}_{\mathcal{S}_o^{(l)}}\big(\Delta_o^{(l)}(x)\big)\right\|_2^2\right]}{\mathbb{E}_{x\sim\bigcup_{o'\neq o} o'}\left[\left\|\text{proj}_{\mathcal{S}_o^{(l)}}\big(\Delta(x)\big)\right\|_2^2\right]} \tag{4}$$

A ratio $\rho \gg 1$ indicates that on-target deltas project more strongly onto $\mathcal{S}_o^{(l)}$ than off-target deltas do, indicating a dedicated subspace. Statistical significance is assessed via a one-sided permutation test; full details are in App. C.2.

**Classification.** $\Delta_o$ are projected into operation-specific subspaces and used to train a 6-way logistic regression classifier with 5-fold stratified cross-validation and random-subspace controls. Silhouette scores on t-SNE projections (van der Maaten and Hinton, 2008) are reported only as a supplementary diagnostic.

### 3.4 Causal Intervention

To evaluate whether operation subspaces are causally implicated in prediction (**RQ2**) and to characterise forward and inverse intervention geometry, targeted activation steering is applied (Fig. 2, **RQ2**). Two direction types are derived from $\Delta_o^{(l)}$: (i) the *principal variance direction*, the top right singular vector of $\mathbf{D}_o^{(l)}$; and (ii) the *mean-shift direction*:

$$\mathbf{v}_o^{(l)} = \frac{\mathbb{E}_i[\Delta_o^{(l)}(x_i)]}{\left\|\mathbb{E}_i[\Delta_o^{(l)}(x_i)]\right\|_2} \tag{5}$$

The latter corresponds to contrastive activation addition (CAA) (Turner et al., 2023; Rimsky et al., 2024). The residual stream is modified as:

$$\mathbf{h}^{(l)}(x) \leftarrow \mathbf{h}^{(l)}(x) + \alpha\, \mathbf{v}_o^{(l)} \tag{6}$$

In the *forward regime* ($+\alpha$), $\mathbf{v}_o^{(l)}$ is applied to clean inputs to shift predictions toward the label associated with $T_o$. In the *inverse regime* ($-\alpha$), it is applied to transformed inputs to restore the clean prediction. The primary metric is the *flip-to-target rate*: the fraction of examples for which the predicted label changes to the target label induced by $T_o$. Evaluation is conducted at $\alpha \in \{5, 10, 20\}$ across all layers, with the best-performing $\alpha$ selected per layer. Results are compared against random steering directions matched in norm.

### 3.5 Mechanistic Localisation

To identify where transformation structure emerges and influences prediction (**RQ4**), two attribution analyses are applied.

**Layer-wise logit decoding.** At each layer $l$, the residual-stream representation is decoded using the unembedding matrix $W_u$ (Belrose et al., 2023):

$$\text{logit_lens}(l) = W_u \cdot \mathbf{h}^{(l)}_{\texttt{resid}} \tag{7}$$

The logit difference $\text{logit}(\text{target}) - \text{logit}(\text{corrupt})$ is tracked across layers to identify the *commitment layer*: the depth at which the correct NLI label first consistently dominates.

**Direct logit attribution (DLA).** The final logit difference is decomposed into additive contributions from model components (Elhage et al., 2021; Nostalgebraist, 2020):

$$\text{logit}(\text{target}) - \text{logit}(\text{corrupt}) = \sum_{l=1}^{L} \left( W_u \cdot \mathbf{h}^{(l)}_{\texttt{attn}} + W_u \cdot \mathbf{h}^{(l)}_{\texttt{mlp}} \right) \tag{8}$$

This decomposition identifies which attention and MLP layers contribute most strongly to the NLI decision, and whether these align with layers exhibiting high subspace selectivity.

## 4 Experimental Setup

**Models.** We evaluate four pretrained decoder-only transformers spanning three model families and the 1.7-4B parameter scale: Qwen3-1.7B (Yang et al., 2025a), Phi-3-mini-4k-instruct (Abdin, 2024), Qwen2.5-3B-Instruct (Yang et al., 2025b), and Gemma3-4B-Instruct (Team et al., 2025) (hereafter Qwen-1.7B, Phi-3, Qwen-3B, and Gemma-4B). No task-specific fine-tuning is applied. Model selection criteria and architectural details are provided in App. C.6.

**Data.** For each operation in $\mathcal{O}$, we generate 300 paired NLI examples $(x_i, T_o(x_i))$ using Claude Sonnet 4.5 (Anthropic, 2025) with structured prompts, operation-specific constraints, and embedded validation checks. Each example starts from a clean premise-hypothesis; the hypothesis is then modified by a target operation while the premise is unchanged. We enforce paired isolation by requiring the modified hypothesis to change only the intended semantic relation, remain grammatical and plausible, and receive the operation-specific target label. Generation was performed across six thematic domains, producing 1,800 contrastive pairs in total. Full generation and validation criteria are provided in App. B.

**Activation extraction and scoring.** The component activations in §3.1 correspond to the TransformerLens hook points `resid_post`, `attn_out`, and `mlp_out`. Model predictions are obtained by scoring next-token probabilities over the NLI label tokens and selecting the highest-probability token.

**Hyperparameters and controls.** The subspace rank is fixed at $k{=}4$, steering interventions are evaluated at $\alpha{\in}\{5, 10, 20\}$; preliminary checks showed qualitatively similar trends for $k{\in}\{2, 8\}$, and classification uses 5-fold stratified cross-validation with disjoint estimation and evaluation splits. Subspace, steering, and classification metrics are compared against matched random-baseline controls. Random subspace controls use rank-$k$ orthonormal bases sampled via QR orthogonalisation of Gaussian matrices, and random steering controls use norm-matched random directions. Additional implementation details and sensitivity checks are provided in App. C.4.

## 5 Results and Analysis

**RQ1 (Structure): Transformation-level effects form partially separable subspaces.** If models distinguish between operations that produce the same entailment label, operation-specific activation changes should exhibit geometric consistency and separability. We first evaluate this structure using subspace selectivity. Across all model-component configurations, best-layer selectivity is significant

Table 2: **Peak (best-layer) subspace selectivity per model and component.** $\rho$**: selectivity ratio.** $d$: Cohen's $d$. $p$: one-sided permutation $p$-value (null distribution constructed from 1000 random orthonormal subspaces). Rel. depth: relative layer position of the peak (0=input, 1=final layer).

| Model | Component | $\rho$ | $d$ | $p$ | Rel. depth |
|---|---|---|---|---|---|
| Qwen-1.7B | resid | 1.34 | 0.54 | 0.002 | 0.93 |
| | attn | 1.51 | 0.66 | 0.001 | 0.67 |
| | mlp | 1.41 | 0.61 | 0.002 | 0.59 |
| Phi-3 | resid | 1.16 | 0.35 | 0.002 | 0.97 |
| | attn | 1.52 | 1.11 | 0.001 | 0.90 |
| | mlp | 1.90 | 1.23 | 0.005 | 0.77 |
| Qwen-3B | resid | 1.26 | 0.46 | 0.003 | 0.77 |
| | attn | 1.30 | 0.57 | 0.001 | 0.69 |
| | mlp | 1.27 | 0.45 | 0.004 | 0.77 |
| Gemma-4B | resid | 1.35 | 1.09 | 0.001 | 0.94 |
| | attn | 1.70 | 2.15 | 0.001 | 0.94 |
| | mlp | 2.07 | 2.40 | 0.002 | 0.82 |

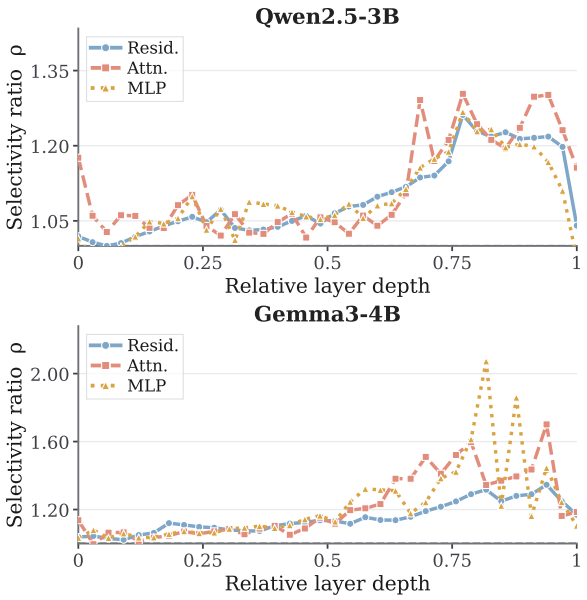


Figure 3: Layer-wise selectivity ratio $\rho$ across relative depth for Qwen-3B and Gemma-4B. Selectivity is low in early layers ($\rho\approx1$) and increases toward the upper network. Full four-model results are in App. D.1.

under permutation testing, with ratios ranging from 1.16 to 2.07× ($p\leq0.009$; Tab. 2). Selectivity is weak in early layers and increases toward the upper network, peaking at relative depths of 0.59-0.97 (Fig. 3). MLP outputs show the strongest selectivity in Phi-3 and Gemma-4B, while attention outputs lead in the Qwen models; the residual stream is weakest across models.

We then evaluate linear separability after projecting deltas into the learned operation subspaces. Six-way classification accuracy reaches 84.8-99.0% at the best layer per model-component combination, substantially above random-subspace baselines (Tab. 8). However, the geometry is not cleanly modular. Per-operation selectivity is heterogeneous: Abstraction has the highest mean selectivity across model-component configurations ($\bar{\rho}$=2.13), followed by Quantification ($\bar{\rho}$=1.96), Co-hyponymy ($\bar{\rho}$=1.88), and Specification ($\bar{\rho}$=1.21). Negation and Antonymy fall below $\rho$=1 in several configurations, indicating overlap with other operation subspaces. This suggests that label-aligned structure partly interacts with operation-specific information; *transformation effects are not reducible to label-level information, but neither are they fully independent operation modules.* Full per-model classification results are reported in App. D.2; per-operation selectivity results are in App. D.

**RQ2 (Causality): Transformation-derived directions can be causally active, but geometric structure does not guarantee steerability.** We investigate whether transformation-induced directions affect predictions under activation steering, comparing CAA and SVD directions (§3.4). Activation steering changes predictions in three of the four models, indicating that transformation-level directions can have causal influence on model outputs. However, this causal effect is not directly predicted by geometric selectivity: Gemma-4B has the strongest peak selectivity in RQ1 ($\rho$=2.07), but exhibits near-zero flip-to-target rates under both CAA and SVD steering. As RQ4 shows, Gemma-4B also exhibits an MLP-dominant, three-phase logit trajectory, suggesting that its operation-level structure may be geometrically organised but not easily controlled by single-direction residual-stream steering. Thus, *geometric separability and causal accessibility should be treated as distinct properties.*

*The divergence between CAA and SVD directions reinforces this point.* Their cosine similarity varies across operations, from 0.459 for Quantification to 0.810 for Specification, suggesting that they capture different aspects of the transformation effect. CAA outperforms SVD in 20 of 33 non-tied comparisons, with a non-tied win rate of 0.606 and a pooled win rate of 0.556 when ties are included (Tab. 3). This advantage is not uniform across models or regimes: Phi-3 favours CAA in both regimes, Qwen-3B is balanced, and Qwen-1.7B reverses across regimes, with CAA leading in the forward setting but SVD outperforming CAA on all six operations in the inverse setting. At the pooled

Table 3: CAA vs SVD steering comparison. Each row pools over operations within a model and direction regime. *Win rate*: proportion of operations for which CAA achieves higher flip-to-target rate than SVD. $\bar{\Delta}$: mean pair difference in flip-to-target rate (positive values favour CAA). $p$: one-sided paired permutation $p$-value (10,000 permutations; App. C.3). The pooled row combines all three models and both regimes.

| Model | Regime | Win rate | $\bar{\Delta}$ | $p$ |
|---|---|---|---|---|
| Qwen-1.7B | forward | 0.667 | +0.160 | 0.062 |
| | inverse | 0.000 | −0.337 | 1.000 |
| Qwen-3B | forward | 0.500 | +0.113 | 0.304 |
| | inverse | 0.500 | +0.023 | 0.424 |
| Phi-3 | forward | 0.833 | +0.210 | 0.062 |
| | inverse | 0.833 | +0.223 | 0.033 |
| **Pooled (3 models, both regimes)** | | **0.556** | **+0.066** | **0.157** |

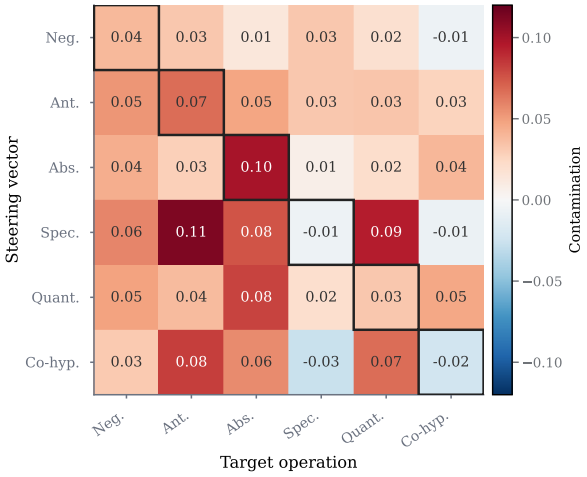


Figure 4: Cross-operation steering contamination. Each cell gives the excess flip-to-target rate when an SVD direction derived from a source operation (rows) is applied to inputs from a target operation (columns), relative to a matched random-direction baseline, averaged over models and layers at $\alpha$=20. On-diagonal cells index within-operation steering; off-diagonal cells index cross-operation interference.

level, the CAA advantage is not statistically significant ($\bar{\Delta}$=0.066, $p$=0.157). A sensitivity analysis at $\alpha \in \{0.5, 1, 2\}$ shows that steering effects decrease substantially at smaller intervention magnitudes (App. D.6).

**RQ3 (Dissociation): Cross-operation interference is sparse but structured, and selectivity does not imply independence.** We measure cross-operation interference by applying an SVD direction derived from one source operation to examples from each target operation. This tests whether steering effects remain confined to the source operation or transfer to other transformations. Off-target effects are generally sparse: at $\alpha$=20, the off-diagonal contamination values are right-skewed, with most pairs clustered near zero (Fig. 4). The five largest off-diagonal values are Specification → Antonymy (0.112), Specification → Quantification (0.095), Co-hyponymy → Antonymy (0.083), Quantification → Abstraction (0.081), and Specification → Abstraction (0.076); all other off-diagonal pairs are at most 0.07. Thus, interference is not uniformly distributed, but concentrated among a small number of semantically adjacent or label-related operation pairs. Per-model and inverse-regime results are reported in App. D.8.

The dissociation between selectivity and independence runs in both directions. Specification has moderate mean selectivity ($\bar{\rho}$=1.21) but is the strongest source of cross-operation interference: Specification → Antonymy (0.112) and Specification → Quantification (0.095) are the two largest off-diagonal values, and Specification appears as source in three of the five largest pairs. Antonymy shows a different pattern: althoguh it is not the main source of interference, it is frequently affected by other operations, with two of the five largest off-diagonal values targeting it. This pattern is consistent with its low subspace selectivity and overlap with other operation subspaces. Abstraction, which has the highest selectivity ($\bar{\rho}$=2.13), has a substantially lower source-side mean (0.028). Negation has both weak selectivity ($\bar{\rho}$=1.03) and the lowest source-side interference (0.018), consistent with a diffuse rather than cleanly isolated signal. Per-model patterns are in App. D.8; interference is broadest for Qwen-3B and most concentrated for Qwen-1.7B, while Gemma-4B shows near-zero contamination. The results indicate that *subspace concentration is neither necessary nor sufficient for intervention-level independence; separability, selectivity, source-side interference, and target-side susceptibility should be treated as distinct mechanistic properties.*

**RQ4 (Localisation): Transformation-dependent label evidence is amplified in mid-to-late layers, with architecture-dependent component pathways.** Layer-wise logit decoding shows that label evidence generally emerges in mid-to-late layers and is amplified in the upper network (Fig. 5; App. D.4). Qwen-1.7B and Phi-3 remain near zero until relative depth 0.6-0.7, while Qwen-3B aligns

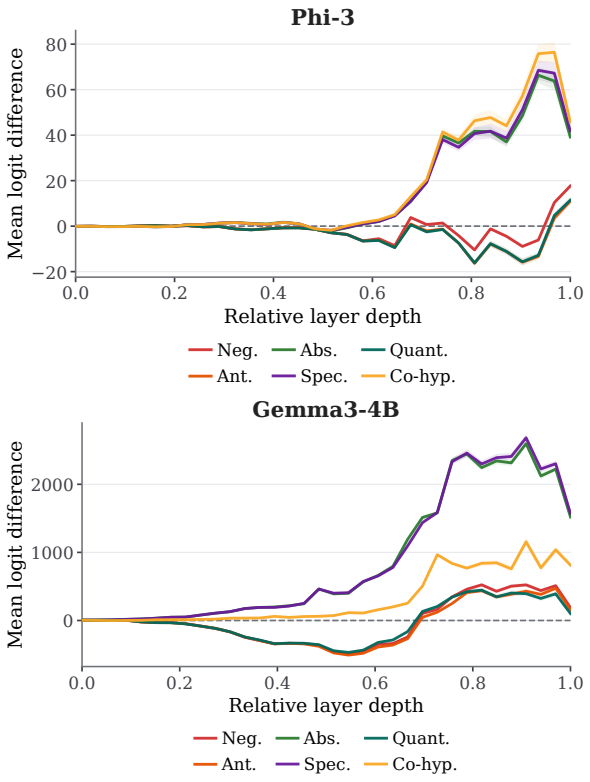


Figure 5: Sample layer-wise logit difference trajectories for Phi-3 and Gemma-4B. Phi-3 shows near-zero evidence until relative depth 0.6-0.7, a brief suppression phase around 0.70-0.75, then sharp rise. Gemma-4B exhibits a three-phase pattern: gradual early rise, mid-network suppression, and strong late amplification with substantially larger magnitudes.

later, rising sharply in the final quarter of the network. Across models, peak logit differences occur at relative depths of 0.86-0.94. Gemma-4B follows a distinct three-phase trajectory: gradual early rise, mid-network suppression around relative depth 0.6-0.65, and strong late amplification at 0.75-0.85.

Direct logit attribution shows that component pathways vary by model family (Fig. 12; App. D.5). Phi-3 is primarily attention-driven, Gemma-4B is MLP-dominant for most operations, Qwen-1.7B is operation-dependent, and Qwen-3B is comparatively balanced. No consistent ordering is observed in the depth at which operations first favour the target label. *These results identify a shared representational phenomenon, but not a shared circuit: transformation-level structure appears across models, while its computational implementation remains architecture-dependent.*

**Overall Interpretation.** The results indicate that transformation-level effects are geometrically organised in model representations, but neither label-independent nor operation-modular. Models do not appear to encode only the final NLI label: transformation effects are linearly decodable, show significant subspace selectivity, and can influence predictions under steering. At the same time, these operations are not represented as cleanly separate modules: their subspaces overlap, steering effects can transfer across operations, and causal accessibility varies across models and regimes. Transformation-level effects are therefore structured, partially shared, and architecture-dependent. Label-level analyses undercharacterise this structure, while stronger circuit-level claims require further localisation beyond the subspace and intervention evidence provided here.

## 6 Conclusion

This paper investigated whether transformer models encode transformation-level effects as structured geometric objects in activation space, distinct from the NLI labels they produce. Using controlled contrastive premise-hypothesis pairs, we measured transformation-induced activation differences, analysed their subspace structure, and tested their causal effect through activation steering.

The results indicate that transformation-level effects are geometrically organised in model representations, but neither label-independent nor operation-modular. Models do not appear to encode only the final NLI label: transformation effects are linearly decodable, show significant subspace selectivity, and can influence predictions under steering. At the same time, these operations are not represented as cleanly separate modules: their subspaces overlap, steering effects can transfer across operations, and causal accessibility varies across models and regimes. Transformation-level effects are therefore structured, partially shared, and architecture-dependent.

These findings constrain mechanistic accounts of semantic inference in transformer models. Label-level analyses undercharacterise this structure, while stronger circuit-level claims require further localisation beyond the subspace and intervention evidence provided here.

## Limitations

This analysis is restricted to decoder-only models at the 1.7-4B scale; generalisation to other architectures, larger models, or explicitly super-

vised NLI systems remains an open question. The paired-isolation assumption is enforced through controlled data construction but cannot be perfectly realised; residual confounds may remain despite random-baseline controls. For example, operations are partially aligned with NLI labels in the controlled dataset, which creates a genuine identifiability concern. We address this with within-label prediction, label-subspace residualisation, and label-aware contamination controls (App. D.9), but fully crossing every operation with every NLI label remains future work. The mechanistic characterisations of semantic operations could in principle be used to construct adversarial inputs that exploit operation-specific directions. Finally, the framework identifies geometric structure and causal relevance but does not localise the underlying circuits. Connecting subspace-level findings to specific mechanistic components remains an important direction for future work.

## Ethical considerations

This work investigates whether semantic operations in LLMs are represented as structured and causally active geometric effects in activation space. Improving mechanistic understanding of semantic inference may support transparency, interpretability, and safety research by enabling more precise analysis of how models process data and produce inference decisions. On the other hand, identifying operation-specific directions and steerable subspaces introduces potential misuse risks. For instance, representations could be exploited to design targeted activation interventions, adversarial prompts, or inference manipulations that selectively alter model behaviour while remaining difficult to detect through output-level analysis.

## References


Marah I. Abdin. 2024. Phi-3 technical report: A highly capable language model locally on your phone. Technical Report MSR-TR-2024-12, Microsoft.

Nura Aljaafari, Danilo Carvalho, and Andre Freitas. 2026. Where do LLMs compose meaning? a layerwise analysis of compositional robustness. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4622–4646, Rabat, Morocco. Association for Computational Linguistics.

Anthropic. 2025. Claude claude-sonnet-4-5. `https://www.anthropic.com`. Large language model used for dataset generation.

Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.

Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. 2023. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. 2023. Discovering latent knowledge in language models without supervision. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Arthur Conmy, Augustine Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. 2023. Towards automated circuit discovery for mechanistic interpretability. In *Advances in Neural Information Processing Systems*, volume 36, pages 16318–16352. Curran Associates, Inc.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, and 6 others. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*. Https://transformer-circuits.pub/2021/framework/index.html.

Matthew Finlayson, Aaron Mueller, Sebastian Gehrmann, Stuart Shieber, Tal Linzen, and Yonatan Belinkov. 2021. Causal analysis of syntactic agreement mechanisms in neural language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1828–1843, Online. Association for Computational Linguistics.

Gottlob Frege. 1892. Über Sinn und Bedeutung. *Zeitschrift für Philosophie und Philosophische Kritik*, 100(1):25–50.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. 2021. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems*, volume 34, pages 9574–9586. Curran Associates, Inc.

Atticus Geiger, Kyle Richardson, and Christopher Potts. 2020. Neural natural language inference models partially embed theories of lexical entailment and negation. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*,

pages 163–173, Online. Association for Computational Linguistics.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Evan Hernandez, Arnab Sen Sharma, Tal Haklay, Kevin Meng, Martin Wattenberg, Jacob Andreas, Yonatan Belinkov, and David Bau. 2024. Linearity of relation decoding in transformer language models. In *Proceedings of the 2024 International Conference on Learning Representations (ICLR)*.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. 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)*, pages 2733–2743, Hong Kong, China. Association for Computational Linguistics.

John Hewitt and Christopher D. Manning. 2019. 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*, pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Inference-time intervention: Eliciting truthful answers from a language model. In *Thirty-seventh Conference on Neural Information Processing Systems(NeurIPS)*.

Bill MacCartney and Christopher D. Manning. 2009. An extended model of natural logic. In *Proceedings of the Eight International Conference on Computational Semantics*, pages 140–156, Tilburg, The Netherlands. Association for Computational Linguistics.

Samuel Marks and Max Tegmark. 2024. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In *First Conference on Language Modeling*.

R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26 (NeurIPS)*, pages 3111–3119.

Richard Montague. 1970a. English as a formal language. In *Linguaggi nella società e nella tecnica*, pages 189–223. Edizioni di Comunità.

Richard Montague. 1970b. Universal grammar. *Theoria*, 36(3):373–398.

Neel Nanda and Joseph Bloom. 2022. Transformerlens. `https://github.com/TransformerLensOrg/TransformerLens`.

Neel Nanda, Lawrence Chan, Tom Liberum, Jess Smith, and Jacob Steinhardt. 2023. Progress measures for grokking via mechanistic interpretability. In *The Eleventh International Conference on Learning Representations (ICLR)*.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics.

Nostalgebraist. 2020. Interpreting gpt: The logit lens. *Blog Post*.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. 2020. Zoom in: An introduction to circuits. *Distill*. Https://distill.pub/2020/circuits/zoom-in.

Kiho Park, Yo Joong Choe, and Victor Veitch. 2023. The linear representation hypothesis and the geometry of large language models. In *Causal Representation Learning Workshop at NeurIPS 2023*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, and 1 others. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. 2024. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15504–15522, Bangkok, Thailand. Association for Computational Linguistics.

Julia Rozanova, Marco Valentino, Lucas Cordeiro, and André Freitas. 2023. Interventional probing in high dimensions: An NLI case study. In *Findings of the*

*Association for Computational Linguistics: EACL 2023*, pages 2489–2500, Dubrovnik, Croatia. Association for Computational Linguistics.

Julia Rozanova, Marco Valentino, and André Freitas. 2024. Estimating the causal effects of natural logic features in transformer-based NLI models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6319–6329, Torino, Italia. ELRA and ICCL.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Eric Todd, Millicent L. Li, Arnab Sen Sharma, Aaron Mueller, Byron C. Wallace, and David Bau. 2024. Function vectors in large language models. In *International Conference on Learning Representations (ICLR)*.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J Vazquez, Ulisse Mini, and Monte MacDiarmid. 2023. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605.

Elena Voita and Ivan Titov. 2020. Information-theoretic probing with minimum description length. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 183–196, Online. Association for Computational Linguistics.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2023. Interpretability in the wild: A circuit for indirect object identification in GPT-2 small. In *The Eleventh International Conference on Learning Representations (ICLR)*.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Zhengxuan Wu, Atticus Geiger, Thomas Icard, Christopher Potts, and Noah D. Goodman. 2023. Interpretability at scale: Identifying causal mechanisms in Alpaca. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025a. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 23 others. 2025b. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Weixiang Zhao, Jiahe Guo, Yang Deng, Tongtong Wu, Wenxuan Zhang, Yulin Hu, Xingyu Sui, Yanyan Zhao, Wanxiang Che, Bing Qin, and 1 others. 2026. When less language is more: Language-reasoning disentanglement makes llms better multilingual reasoners. *Advances in Neural Information Processing Systems*, 38:38608–38642.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, and 2 others. 2023. Representation engineering: A top-down approach to AI transparency. *Preprint*, arXiv:2310.01405.

# A Linguistic and Logical Foundations

## A.1 Sentence Meaning as Functional Constraint

Formal semantics identifies the meaning of a sentence with a function of the meanings of its parts (Frege, 1892). Under this compositional view, $[\![S]\!]$ is determined by the meanings of $S$'s constituent expressions $s_1, s_2, \ldots, s_n$ with their syntactic combination (Montague, 1970b,a):

$$[\![S]\!] \;=\; f([\![s_1]\!],\; [\![s_2]\!],\; \ldots,\; [\![s_n]\!])\,, \qquad (9)$$

so that to know what $S$ means is to know how its parts contribute to the whole. This determines which worlds $S$ permits and which it excludes, and sentence-pair relations follow directly: $P$ entails $H$ when $[\![P]\!]^{-1}(1) \subseteq [\![H]\!]^{-1}(1)$, $P$ contradicts $H$ when the two sets are disjoint, and the pair is neutral otherwise.

Natural language inference (NLI) provides a task-level approximation to this semantic picture (Bowman et al., 2015; Williams et al., 2018). Given a premise $P$ and a hypothesis $H$, the problem is to determine which of these three relations holds; a model that solves it correctly must therefore be sensitive, at some level, to the compositional meaning constraints that determine the answer. This makes NLI a principled testbed for studying how models represent sentence meaning: not only through

their final labels, but through the relation the labels reflect.

### A.2 Semantic Operations as Local Meaning Changes

A semantic operation acts locally: it modifies a targeted expression or semantic part within the hypothesis while keeping the premise, and as much of the remaining hypothesis as possible fixed. Under the compositional view of Eq. 9, this local change can alter the meaning of the whole sentence and the semantic relation between premise and hypothesis. This link between constituent-level change and sentence-level consequence makes operations an appropriate unit of analysis for studying *how meaning changes are encoded in model representations*.

Natural logic formalises this idea through a small set of semantic relations between expressions: forward entailment ($\sqsubset$), reverse entailment ($\sqsupset$), equivalence ($\equiv$), negation ($\wedge$), alternation ($|$), cover ($\smile$), and independence ($\#$) (MacCartney and Manning, 2009). These local relations interact with the related context, where that changing a word, an operator, or the level of specificity can alter the relation between the full premise and the hypothesis. For example, replacing a term with a more specific expression makes the hypothesis stronger than the premise supports ($\sqsubset$). In each case, the edit is local but its inferential effect is sentence-wide.

The six operations studied in this paper instantiate this general pattern across several relation types: Negation and Quantification modify logical structure; Antonymy and Co-hyponymy modify lexical relations; and Abstraction and Specification modify semantic granularity. They provide a controlled set of meaning-altering transformations, allowing us to ask whether models distinguish the operation that changes the relation from the label it produces.

### A.3 Atomicity and Paired Isolation

The operations defined above are local by design, but locality alone does not guarantee that an activation difference reflects only the intended semantic change. We use *atomicity* in an operational sense: within a constructed pair, the transformed hypothesis should differ from the clean hypothesis along one primary semantic dimension. A SPECIFICATION pair, for instance, should add a more specific constraint without introducing an unrelated contradiction.

This motivates *paired isolation*. Each example consists of a clean input $(P, H_{\text{clean}})$ and a transformed input $(P, H_o)$, where the premise is fixed and the hypothesis is modified by a target operation $o$. Under atomicity, the activation difference can be defined as:

$$\Delta_o^{(l)}(x_i) = \underbrace{\boldsymbol{\delta}_o^{(l)}}_{\text{operation effect}} + \underbrace{\boldsymbol{\varepsilon}_i^{(l)}}_{\text{instance noise}}, \quad \mathbb{E}_i\left[\boldsymbol{\varepsilon}_i^{(l)}\right] \approx \mathbf{0}, \tag{10}$$

where $\boldsymbol{\delta}_o^{(l)}$ is the shared operation-induced activation difference and $\boldsymbol{\varepsilon}_i^{(l)}$ is zero-mean instance noise. The activation difference is intended to capture the representational change associated with $T_o$, rather than random variation between unrelated parts.

The assumption is approximate. A negation marker may also alter syntactic structure; an antonym may differ from the original predicate in frequency or register; and a specification may introduce additional lexical material. The role of controlled generation and validation is to make the target operation the dominant source of variation in $\Delta_o^{(l)}$, not to guarantee a perfectly isolated semantic atom. The main analyses combine several checks: random-subspace controls test whether the observed structure exceeds arbitrary low-dimensional projections; label-controlled analyses address the largest systematic confound; and cross-operation contamination tests whether operation-derived directions remain specific or transfer to other transformations.

**Connection to the mechanistic analysis.** A model whose internal representations are sensitive to sentence meaning should, under this view, respond systematically when a semantic operation is applied. The compositional structure of Eq. 9 motivates the central test: whether a local change to the hypothesis, with the premise and surrounding context held fixed, leads to a consistent activation difference in the model's representation of the full input. This activation difference, $\Delta_o^{(l)}(x_i)$, is the quantity our mechanistic analysis is designed to detect. Whether it is structured, operation-specific, and causally active in the model's predictions is the central question the paper inveistage.

## B Data Generation Details

This section describes the generation procedure for the contrastive NLI pairs used in the main experiments. The dataset was an English dataset, generated using Claude Sonnet 4.5 (Anthropic,

2025) under a structured prompt with operation-specific constraints. The goal was to produce paired examples in which the clean pair expressed entailment and the modified pair differed from it by exactly one controlled semantic transformation. As the main analysed quantity is $\Delta_o^{(l)} = \mathbf{h}^{(l)}(P, H_o) - \mathbf{h}^{(l)}(P, H_{\text{clean}})$, the dataset must minimise all variation between the clean and modified hypotheses except the target operation; this motivates the controlled design described below.

### B.1 Operation Selection Criteria

The six operations were selected to provide coverage of the primary lexical semantic relation types identified in natural logic accounts of NLI (MacCartney and Manning, 2009), while satisfying the methodological requirements of the mechanistic analysis.

**Grounding in natural logic.** The six operations instantiate distinct lexical relation types within the natural logic framework of MacCartney and Manning (2009) (see App. A.2 for the formal grounding). NEGATION generates the exhaustive exclusion relation ($\wedge$); ANTONYMY and CO-HYPONYMY both instantiate alternation ($|$) but are distinguished here by structural and representational criteria: antonymy involves direct semantic opposition, whereas co-hyponymy involves category sibling substitution under a shared hypernym. QUANTIFICATION generates exclusion via quantifier substitution (e.g. *some* $\rightarrow$ *no*). ABSTRACTION and SPECIFICATION instantiate reverse and forward entailment ($\sqsupset$, $\sqsubset$) via hypernymy-axis substitution, producing NEUTRAL labels.

**Selection criteria.** Three criteria governed the final selection. First, each operation must modify the hypothesis along a single semantic dimension, satisfying the paired-isolation requirement of the mechanistic analysis (§3.1). This excludes operations requiring simultaneous modification of multiple constituents or world-knowledge dependencies that cannot be controlled lexically. Second, the set must cover both non-entailment label outcomes (CONTRADICTION and NEUTRAL), which is necessary for the label-controlled analyses in Appendix D.9. Third, operations should be formally and structurally distinct enough to test whether transformation-level structure generalises across qualitatively different types of meaning-altering edits.

**Excluded operations.** Meronymy (part-whole substitution) was piloted but produced inconsistent label assignments due to pragmatic variability in part-whole entailment relations, violating the first criterion. Scalar implicature was excluded because label assignment depends on pragmatic context rather than lexical substitution alone. Presupposition-altering operations and causal or temporal modifications were excluded because they require multi-constituent changes or introduce uncontrollable world-knowledge dependencies. The six retained operations represent the maximal set satisfying all three criteria under the controlled generation protocol described in §B.3.

### B.2 Sentence Universe

Figure 6 provides an overview of the controlled sentence universe. The universe consists of declarative English sentences describing plausible everyday events, states, or entity relations. Premises are designed to have a clean hypothesis by direct lexical entailment, without requiring inference chains, background knowledge, or world-knowledge dependencies. Modified hypotheses differ from the clean hypothesis by a semantic operation.

**Sentence length and complexity.** Table 5 reports descriptive statistics for the generated pairs. Premises range from 7 to 18 tokens (median 10), clean hypotheses from 2 to 15 tokens (median 8), and modified hypotheses from 2 to 16 tokens (median 9) across all operations. The near-identical length distributions of $H_{\text{clean}}$ and $H_o$ confirm that transformations do not systematically alter sentence length, consistent with the paired-isolation design. QUANTIFICATION produces the longest sentences (premise median 12, hypothesis median 10–11), reflecting the need to include explicit quantifier expressions. ANTONYMY produces the shortest hypotheses (median 4 tokens), consistent with the targeted single-word substitution it requires.

### B.3 Prompting Procedure

The generator model received a structured prompt organised into five components: (i) an operation definition specifying the semantic transformation in formal terms; (ii) a set of word pair examples illustrating valid substitutions for that operation; (iii) explicit label flip rules specifying the required baseline and target labels; (iv) a validation checklist that the model was required to apply to every example before inclusion; and (v) an output

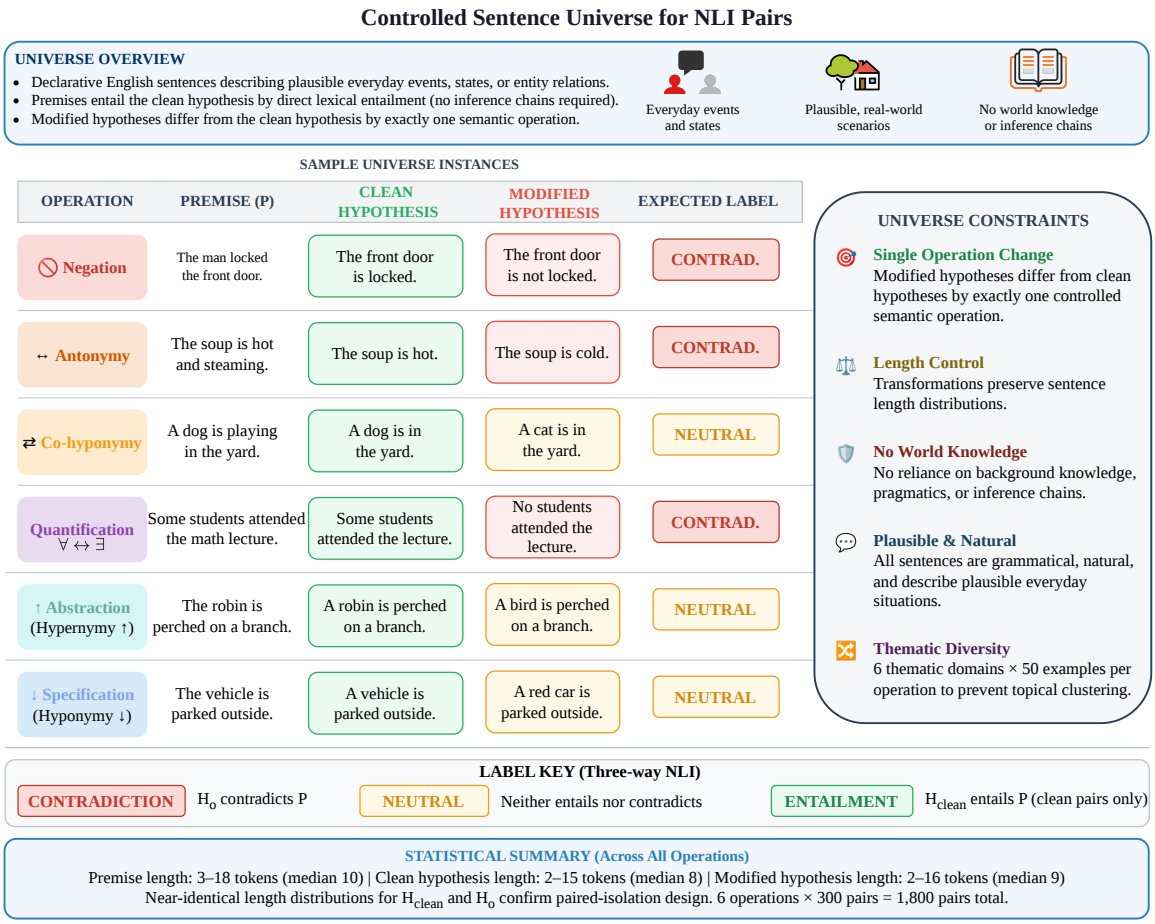


Figure 6: Controlled sentence universe for contrastive NLI pair generation. Each pair consists of a fixed premise $P$, a clean hypothesis $H_{\text{clean}}$ (labelled ENTAILMENT), and a modified hypothesis $H_o = T_o(H_{\text{clean}})$ differing from $H_{\text{clean}}$ by exactly one semantic operation. One representative example is shown per operation.

format specification requiring structured JSON with fields for premise, clean hypothesis, modified hypothesis, baseline label, target label, operation type, original span, and modified span. Table 4 summarises the prompt components and their functions.

Table 4: Prompt structure used for controlled data generation.

| Component | Function |
|---|---|
| Operation definition | Defines the semantic edit and target label transition in formal terms. |
| Substitution examples | Provides valid word-pair examples for the target operation. |
| Label flip rules | Specifies baseline ENTAILMENT and target label explicitly. |
| Validation checklist | Lists conditions every example must satisfy before inclusion. |
| Output schema | Requires structured JSON with fixed fields for downstream filtering. |
| Regeneration rule | Instructs the model to discard and replace any failing example. |

**Operation definitions in the prompt.** Each operation was defined in semantic rather than surface-form terms, with explicit label flip rules:

- **Negation**: insert a negation marker into the main predicate of the hypothesis; baseline ENTAILMENT → target CONTRADICTION.
- **Antonymy**: replace a content word with a direct, non-gradable antonym; baseline ENTAILMENT → target CONTRADICTION.
- **Quantification**: replace a quantifier with a maximally distinct alternative; baseline ENTAILMENT → target CONTRADICTION.
- **Abstraction**: replace a specific term with a superordinate hypernym; baseline ENTAILMENT → target NEUTRAL.
- **Specification**: replace a general term with a more specific hyponym; baseline ENTAILMENT → tar-

get NEUTRAL.

- **Co-hyponymy**: replace a term with a sibling category member sharing the same hypernym; baseline ENTAILMENT → target NEUTRAL.

**Validation requirements embedded in the prompt.** The prompt required the generator to verify the following conditions for every example before inclusion:

1. Both clean and modified hypotheses are grammatical and natural English sentences.
2. Both describe plausible real-world situations.
3. The clean hypothesis is unambiguously entailed by the premise.
4. The modified hypothesis bears the target non-entailment label.
5. The modified hypothesis does not accidentally remain entailed by the premise.
6. The transformation involves exactly one semantic change, with no additional lexical or syntactic modifications if not necessary.

**Self-filtering.** Generation included an explicit self-check stage: the model was instructed to discard and regenerate any example failing any of the above conditions. Low-margin examples, those for which a competent reader might hesitate before assigning the intended label, were identified as a major failure mode in pilot generations and were explicitly prohibited.

**Scale and batching.** Each operation was generated in six thematic batches of 50 examples, yielding 300 examples per operation and 1,800 pairs in total across six operations. Each operation was generated independently and saved as a separate file.

**Output format.** The prompt required JSON output. A representative example for NEGATION is shown below:

```
{
  "id": "neg_0042",
  "premise":  "The company hired three new
               engineers last month.",
  "clean":    "The company hired new engineers.",
  "corrupted": "The company did not hire
               new engineers.",
  "baseline_label": "entailment",
  "modified_label": "contradiction",
  "operation":      "NEGATION",
  "original_span":  "hired",
  "modified_span":  "did not hire"
}
```

### B.4 Filtering and Validation

After generation, files were inspected programmatically to ensure schema consistency. Each JSON object was validated to contain the required fields: `id`, `premise`, `clean`, `corrupted`, `baseline_label`, `modified_label`, `operation`, `original_span`, and `modified_span`.

### B.5 Relation to the Main Experiments

The purpose of the synthetic generation procedure was not to approximate the full distribution of naturally occurring NLI examples, but to construct minimally varying contrastive pairs suitable for mechanistic analysis. The controlled procedure therefore prioritises paired isolation and label clarity over ecological breadth.

Since the data are generated by LLMs under explicit structural constraints, residual artefacts may remain, including lexical regularities or templatic biases weaker in naturally occurring corpora. The random-subspace controls, cross-model comparisons, and cross-operation analyses help to bound this risk, but do not eliminate it. The results should therefore be interpreted as evidence about transformation-level structure under controlled contrastive conditions, rather than as a claim about the full natural distribution of NLI examples.

Table 5: Descriptive statistics of the generated sentence universe. Values reported as median (min–max) per field and operation.

| Operation | Premise | $H_{\text{clean}}$ | $H_o$ |
|---|---|---|---|
| ABSTRACTION | 11 (7–15) | 9 (5–13) | 11 (8–15) |
| ANTONYMY | 10 (7–15) | 5 (4–6) | 5 (4–6) |
| CO HYPONYMS | 8 (4–13) | 8 (4–12) | 8 (4–13) |
| NEGATION | 10 (7–14) | 6 (3–13) | 8 (4–14) |
| QUANTIFICATION | 12 (8–18) | 10 (5–15) | 11 (6–16) |
| SPECIFICATION | 8 (7–13) | 6 (4–11) | 6 (4–11) |
| *All operations* | 10 (4–18) | 8 (4–15) | 9 (4–16) |

# C Evaluation Metrics, Controls, and Implementation Details

## C.1 Criteria for Transformation-Based Reasoning

We operationalise transformation-based reasoning as the existence of structured, reusable directional effects in representation space. A model exhibits transformation-based reasoning with respect to operation $o$ if four criteria are satisfied. **(i) Consistency**: the transformation induces a stable subspace $\mathcal{S}_o^{(l)}$ that generalises across input instances. **(ii) Separability**: distinct operations occupy distinguishable regions of representation space, such that a linear classifier separates them substantially above chance. **(iii) Causality**: the transformation can be induced or reversed via targeted intervention along directions in $\mathcal{S}_o^{(l)}$, producing the corresponding shift in model output. **(iv) Independence**: intervening along $\mathbf{v}_o$ does not systematically shift model predictions toward the label of a different operation $o'$, as measured by cross-operation steering contamination. Criteria (i)-(ii) are correlational, establishing that structure exists in representations. Criteria (iii)-(iv) are causal, establishing that the structure is used by the model. The main evaluation (§3, §5) tests each criterion via distinct analyses: selectivity and classification address (i)-(ii); activation steering addresses (iii); cross-operation interference addresses (iv).

We analyse the displacement distribution $\Delta_o^{(l)}$ in two complementary ways. The mean-shift direction captures the average operation-induced displacement and is the direction used in contrastive activation addition (CAA) (Turner et al., 2023; Rimsky et al., 2024). The SVD basis captures the dominant axes of variation around this displacement distribution. This distinction matters because a semantic operation may be expressed partly as a common shift shared across examples and partly as structured variation due to lexical, syntactic, or contextual realisation. Comparing CAA and SVD tests whether causal control is better explained by the average displacement or by a principal direction of operation-specific variation.

This formulation connects representation geometry with causal intervention. Prior work shows that relational and truth-conditional information can be recovered from linear structure in transformer activations (Park et al., 2023; Hernandez et al., 2024; Marks and Tegmark, 2024), while activation-steering work shows that contrastive activation differences can change model behaviour when added to internal states (Turner et al., 2023; Rimsky et al., 2024; Zou et al., 2023). Related subspace-based interventions have also been used to separate behaviourally relevant representational factors (Zhao et al., 2026).

The operation subspace $\mathcal{S}_o^{(l)}$ captures the principal directions along which $T_o$ displaces representations at layer $l$, analogous to a tangent space in differential geometry. The analogy is informal: the underlying representation manifold need not be smooth, and $\mathcal{S}_o^{(l)}$ is a linear approximation to a potentially nonlinear displacement structure.

## C.2 Evaluation Metrics

**Subspace selectivity.** The selectivity ratio $\rho_o^{(l)}$ (Equation 4) measures the relative projection energy of an operation onto its own subspace versus the average projection energy of other operations onto the same subspace. A ratio of $\rho = 1$ indicates no preferential response; values greater than 1 indicate selective response. Formally, for operation $o$ at layer $l$, the selectivity ratio is defined as:

$$\rho_o^{(l)} = \frac{\frac{1}{N_o}\sum_{i=1}^{N_o}\left\|\mathrm{proj}_{\mathcal{S}_o^{(l)}}\big(\Delta_o^{(l)}(x_i)\big)\right\|_2^2}{\frac{1}{\sum_{o'\neq o} N_{o'}}\sum_{o'\neq o}\sum_{j=1}^{N_{o'}}\left\|\mathrm{proj}_{\mathcal{S}_o^{(l)}}\big(\Delta_{o'}^{(l)}(x_j)\big)\right\|_2^2} \tag{11}$$

The numerator is the mean projection magnitude of operation $o$'s deltas onto $\mathcal{S}_o^{(l)}$ (on-target); the denominator pools all deltas from operations $o' \neq o$ onto the same subspace (off-target). Selectivity is reported as a ratio rather than a difference to ensure scale invariance across operations and layers with different overall activation magnitudes.

**Cluster Metrics.** To assess geometric separation independently of classifier choice, we compute the silhouette score on projected activation deltas:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \tag{12}$$

where $a(i)$ is the mean intra-cluster distance and $b(i)$ is the mean nearest-cluster distance for sample $i$. Values range from $-1$ to 1, with higher values indicating better-separated clusters.

**Steering and cross-operation contamination.** For causal intervention experiments (§3.4), we report:

- **Flip-to-target rate**: fraction of examples whose predicted NLI label changes from the pre-intervention label to the target label (the label associated with the applied operation).
- **Contamination/Off-target flip rate**: fraction of examples whose predicted label changes to a label other than the target.

The term *contamination*, used in §5 and Fig. 4, refers to a distinct quantity defined in the cross-operation steering analysis: the excess flip-to-target rate when a source operation's direction is applied to a different target operation, relative to a matched random-direction baseline.

### C.3 Statistical Testing

**Selectivity significance.** Statistical significance is assessed using a one-sided permutation test that evaluates whether the observed on-target projection magnitudes exceed those expected under random subspace alignment. For each configuration, an empirical null distribution is constructed by sampling random orthonormal subspaces of rank $k$ and recomputing the selectivity ratio. Reported $p$-values correspond to the proportion of random samples that yield selectivity ratios greater than or equal to the observed value. Cohen's $d$ is reported as an effect size.

**Classification significance.** We compare SVD-based classification accuracy against the random-baseline distribution using permutation tests with 1000 random-subspace draws. We report the proportion of random draws that exceed the observed accuracy as an empirical $p$-value.

**Paired permutation test for direction comparison.** To compare CAA mean-shift and SVD principal variance directions under activation steering (§5, RQ2), we use a paired sign-flip permutation test. Each unit of comparison is a (model, direction, operation) triple for which a flip-to-target rate is available for both direction types, yielding $n = 36$ paired rows (three models, two steering regimes, six operations; Gemma3-4B excluded as non-informative, §5). The test statistic is the mean pair difference $\bar{\Delta} = \mathrm{mean}(\mathrm{CAA} - \mathrm{SVD})$. Under the null hypothesis of no systematic difference, the sign of each pair difference is randomly flipped across 10,000 permutations. The one-sided $p$-value is the proportion of permuted means that equal or exceed the observed $\bar{\Delta}$.

### C.4 Hyperparameters and Random-Baseline Controls

**Subspace estimation.** The activation difference $\Delta_o^{(l)}(x_i)$ asks a more specific question than standard probing: not only whether an NLI label is decodable from a representation, but whether applying the same semantic operation induces a consistent geometric effect across examples. For each model, component, and layer, activation deltas $\Delta_o^{(l)} \in \mathbb{R}^{N_o \times d}$ are computed as the difference between corrupted and clean activations. The operation subspace $\mathcal{S}_o^{(l)}$ is estimated via truncated SVD applied to the mean-centred delta matrix, retaining the top $k$=4 right singular vectors as an orthonormal basis. This rank is fixed uniformly across all operations, layers, components, and models; sensitivity was verified qualitatively at $k \in \{2, 8\}$, yielding no substantive change in relative ordering. Up to $N$=300 paired examples are used per operation; all layers are extracted at stride 1, covering indices 0 through $L-1$.

Two assumptions underlie this procedure. **Assumption 1: Paired isolation** (see App. A.3 and App. B for the formal statement and the generation protocol that enforces it). **Assumption 2: Low-rank structure**: variability in activation deltas is assumed to be approximately low-rank, with the transformation effect concentrated in a small number of directions relative to $d$. The true intrinsic dimensionality of each operation's representation is unknown and may vary across operations and model families. **Assumption 3: Linearity**: the SVD framework and linear intervention assume that transformation effects are approximately linear in representation space; nonlinear displacements are captured only in projection, potentially underestimating operation-specific structure.

**Random-subspace baselines.** To confirm that SVD-derived subspaces capture operation-specific structure rather than arbitrary projections, each subspace metric is compared against a random-baseline control. Random orthonormal subspaces of rank $k$=4 are generated by drawing a Gaussian matrix $G \in \mathbb{R}^{d \times k}$ with a layer-dependent seed, then applying QR decomposition: $G = QR$, $Q \in \mathbb{R}^{d \times k}$. Five independent random seeds are

evaluated per layer, and the mean classification accuracy across seeds and layers is reported as the chance-level reference.

**Classification.** Linear classifiers are evaluated using 5-fold stratified cross-validation (`random_seed`=42), with estimation and evaluation splits fully disjoint across folds. Three classifiers are compared at each layer: logistic regression ($C$=0.1, `liblinear` solver, balanced class weights, `max_iter`=1000), linear SVM ($C$=0.1, balanced class weights), and RBF-kernel SVM ($C$=1.0). The best-layer accuracy reported for each model-component combination is the maximum mean accuracy across layers for the logistic regression classifier, consistent with the classification results in Table 8.

**Selectivity permutation test.** Statistical significance of the selectivity ratio $\rho_o^{(l)}$ is assessed using a one-sided permutation test over the selectivity matrix. The observed statistic is the mean on-diagonal projection magnitude. Under each of $B$=5,000 permutations, the column indices of the selectivity matrix are randomly shuffled, and the mean of the resulting pseudo-diagonal is recorded as the null distribution. The empirical $p$-value is computed as $(|\{b : \text{null}_b \geq \text{observed}\}| + 1)/(B + 1)$.

**Steering hyperparameters.** Activation steering is applied at all layers using the residual stream component (`resid_post`). The top singular vector of the SVD subspace (index 0) is used as the SVD steering direction; the mean-shift direction (CAA) is computed as the arithmetic mean of the per-example deltas for each operation. Steering magnitude is swept over $\alpha \in \{5, 10, 20\}$ for the main evaluation, and over $\alpha \in \{0.5, 1.0, 2.0\}$ for the sensitivity analysis reported in Appendix D.6.

**CAA vs SVD comparison permutation test.** The significance of the direction-type advantage (CAA over SVD) is assessed using a one-sided paired sign-flip permutation test. For each model and regime, the per-operation differences $d_o = \text{flip_to_target}_{\text{CAA}} - \text{flip_to_target}_{\text{SVD}}$ are computed at the best $\alpha$ per layer. Under each of $B$=10,000 permutations, each $d_o$ is independently multiplied by a random sign $s_o \in \{-1, +1\}$ drawn uniformly, and the mean of the permuted differences is recorded. The $p$-value is $(|\{b : \bar{d}_b \geq \bar{d}_{\text{obs}}\}| + 1)/(B + 1)$, where $\bar{d}_{\text{obs}}$ is the observed mean difference.

**Random-baseline steering directions.** Each structured steering vector (SVD or CAA) is matched against a random unit-norm direction of the same dimensionality, scaled to the same $\ell_2$ norm as the structured vector. Three independent random seeds are used per evaluation to estimate the null-distribution flip rate. The contamination metric for cross-operation interference is defined as the difference between the observed and random-baseline flip-to-other rates: contamination = flip_to_other − rand_flip_to_other, where positive off-diagonal contamination indicates that the structured direction disrupts another operation's predictions more than a random direction of equal magnitude.

### C.5 Additional Analysis Methods

#### C.5.1 Layer-wise emergence

We track selectivity $\rho_o^{(l)}$ and classification accuracy as a function of layer depth $l$ to identify where transformation structure emerges. Full layer-wise profiles are shown in App. D.1 and Fig. 10.

### C.6 Model Specifications

Table 6 summarises the Transformer models used in this study. We select four open-weight decoder-only models spanning the 1.7-4B parameter range: Qwen3-1.7B (Yang et al., 2025a), Qwen2.5-3B-Instruct (Qwen2.5-3B) (Yang et al., 2025b), Phi-3-mini-4k-instruct (Phi-3) (Abdin, 2024), and Gemma3-4B-Instruct (Gemma3-4B) (Team et al., 2025). The selection is governed by three criteria:

- **Architectural comparability.** All four models are decoder-only Transformers while differing in training data composition, tokenisation, and optimisation regimes. This allows us to test whether transformation-level structure is consistent across model families rather than specific to a single implementation.

- **Controlled scale.** The 1.5-4B parameter range provides a regime in which models are sufficiently expressive to exhibit non-trivial semantic behaviour while remaining comparable in depth and width. This reduces confounding effects from large differences in capacity and facilitates layer-wise analysis at matched relative depths.

- **Pretraining distribution.** Qwen models are trained on multilingual corpora, while Phi-3 and Gemma3-4B are predominantly English-focused.

Table 6: Model properties across architectures. Params: number of parameters, Layers: number of layers, $D_{model}$: size of word embeddings and hidden states, Heads: number of attention heads, Act.: Activation function, MLP Dim: dimensionality of the FF layers.

| Model | Params | Layers | $D_{model}$ | Heads | Act. | MLP Dim |
|---|---|---|---|---|---|---|
| gemma-3-4b-it | 3.4B | 34 | 2560 | 8 | gelu_pytorch_tanh | 10240 |
| phi-3 | 3.6B | 32 | 3072 | 32 | SiLU | 8192 |
| Qwen2.5-3b-it | 3.0B | 36 | 2048 | 16 | SiLU | 8192 |
| Qwen3-1.7B | 1.5B | 28 | 2048 | 16 | SiLU | 2048 |

This variation allows us to test whether the emergence of transformation-level structure depends on linguistic diversity in the training data.

All models are used in their released inference configurations and are analysed without fine-tuning. Licences are as follows: Qwen models are released under the Qwen licence agreement, Gemma under the Gemma Terms of Use, and Phi under the MIT licence. All experiments are conducted in accordance with the intended research use of these models.

### C.7 Software and Implementation

Experiments were conducted on H100 GPU, using Python 3.12 with the following core dependencies: NumPy 1.26.4, pandas 2.2.2, PyTorch 2.10.0 (Paszke et al., 2019), TransformerLens 2.18.0 (Nanda and Bloom, 2022), Transformers 4.57.6, scikit-learn 1.7.0 (Pedregosa et al., 2011), SciPy 1.15.3.

## D Additional Results

### D.1 Selectivity

Fig. 7 shows layer-wise selectivity ratio $\rho$ across relative depth for all model-component combinations, while Fig. 8 shows layer-wise selectivity curves for all four models and components. Tab. 7 gives the full numerical results. Fig. 9 shows per-operation selectivity ratios at the peak-selectivity layer for all four models. The heterogeneity described in §5 is consistent across models: Abstraction exhibits the highest per-operation selectivity across most model-component combinations, followed by Quantification and Co-Hyponyms, while Negation and Antonymy fall below $\rho = 1$ in several configurations.

### D.2 Classification

Tab. 8 reports best-layer linear classification accuracy for the six-way operation classification task across all model-component combinations. All

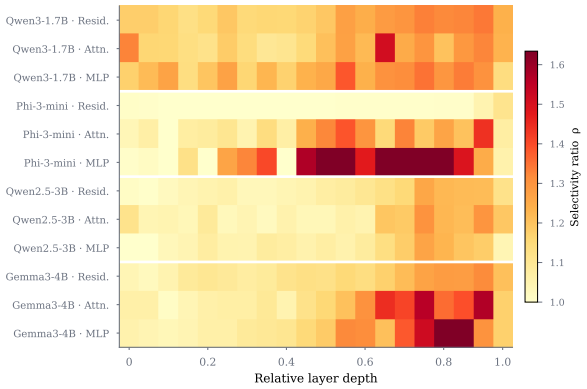


Figure 7: Selectivity ratio across relative layer depth for all model-component combinations. Rows correspond to model-component pairs and columns to relative layer depth. Warmer colours indicate higher selectivity. Selectivity increases toward the upper network across models, with strongest effects in MLP components.

configurations exceed the chance level of $1/6 \approx 16.7\%$ by a substantial margin, with most reaching above $93\%$. The best-layer for classification does not always coincide with the best-layer for selectivity, indicating that the two metrics assess different but complementary aspects of the representational geometry. Fig. 10 shows the full layer-wise accuracy profiles.

### D.3 Clustering

Silhouette scores are computed on t-SNE projections of activation deltas at the best-silhouette layer for each model-component combination (Tab. 9). All twelve scores are negative, indicating that projected representations do not form well-separated clusters in two dimensions. This is consistent with the subspace geometry characterised in §5: the full $d$-dimensional subspaces are linearly separable to a classifier (Tab. 8), but two-dimensional projections do not yield sharp cluster boundaries, reflecting the graded overlap in the selectivity ratio profiles. Silhouette scores are treated as a convergent geometric diagnostic rather than primary evidence.

### D.4 Logit Lens Trajectories

Fig. 11 shows layer-wise logit difference trajectories for all four models. Qwen3-1.7B and Qwen2.5-3B follow the same late-amplification pattern described in §5: near-zero trajectories through early and middle layers, with sharp amplification concentrated in the final 10-15% of layers.

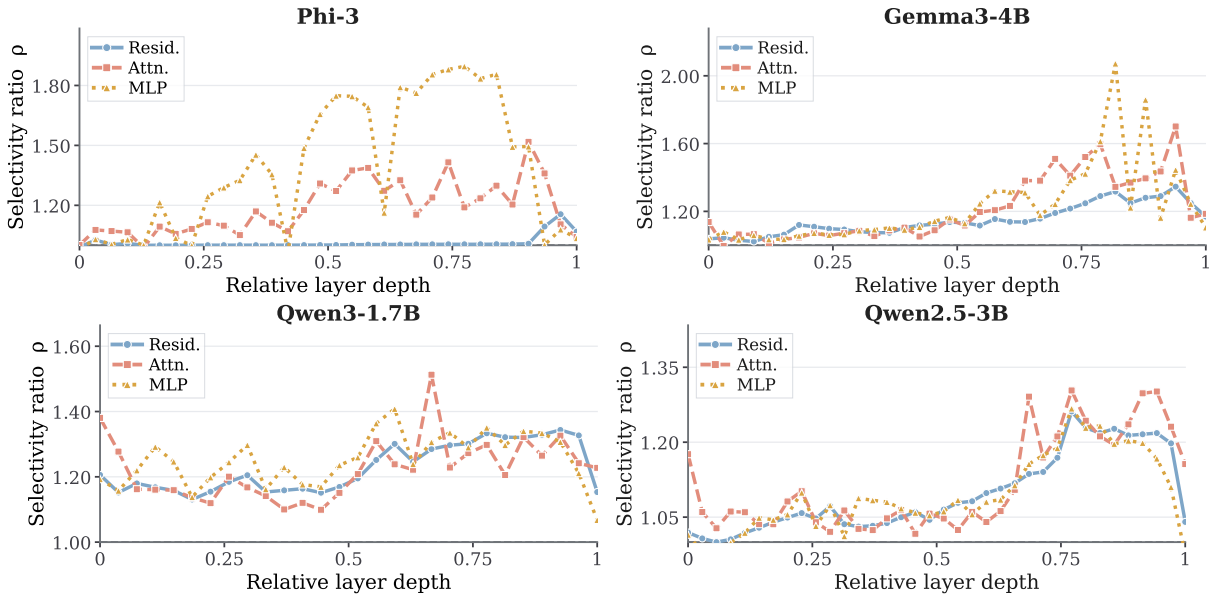


Figure 8: Layer-wise selectivity ratio $\rho$ across relative depth. Selectivity is low in early layers ($\rho \approx 1$) and increases toward the upper network, with strongest effects in MLP components.

Table 7: Full subspace selectivity results. Layer: absolute layer index at which $\rho$ is maximised. Rel.: relative depth (layer / (total layers $-$ 1)). $\mu_{\text{on}}$, $\mu_{\text{off}}$: mean projection magnitudes of on-target and off-target activation deltas onto the operation subspace at the peak layer. All permutation $p$-values computed with 1000 random orthonormal subspace draws from the Gaussian manifold.

| Model | Component | $\rho$ | $d$ | $p$ | Layer | Rel. | $\mu_{\text{on}}$ / $\mu_{\text{off}}$ |
|---|---|---|---|---|---|---|---|
| Qwen3-1.7B | resid_post | 1.344 | 0.537 | 0.002 | 25 | 0.93 | 166.8 / 124.2 |
| | attn_out | 1.513 | 0.658 | 0.001 | 18 | 0.67 | 7.2 / 4.8 |
| | mlp_out | 1.408 | 0.607 | 0.002 | 16 | 0.59 | 9.3 / 6.6 |
| Qwen2.5-3B | resid_post | 1.262 | 0.464 | 0.003 | 27 | 0.77 | 7.3 / 5.8 |
| | attn_out | 1.304 | 0.569 | 0.001 | 27 | 0.77 | 2.3 / 1.8 |
| | mlp_out | 1.267 | 0.451 | 0.004 | 27 | 0.77 | 5.0 / 4.0 |
| Phi-3 | resid_post | 1.155 | 0.353 | 0.002 | 30 | 0.97 | 24.4 / 21.1 |
| | attn_out | 1.518 | 1.111 | 0.001 | 28 | 0.90 | 2.4 / 1.6 |
| | mlp_out | 1.896 | 1.233 | 0.005 | 24 | 0.77 | 1.2 / 0.7 |
| Gemma3-4B | resid_post | 1.346 | 1.095 | 0.002 | 31 | 0.94 | 1879.5 / 1396.5 |
| | attn_out | 1.701 | 2.154 | 0.001 | 31 | 0.94 | 482.1 / 283.4 |
| | mlp_out | **2.072** | **2.402** | 0.002 | 27 | 0.82 | 234.1 / 113.0 |

### D.5 Direct Logit Attribution

Direct logit attribution (DLA) decomposes the final logit difference into additive contributions from attention and MLP components at each layer (§3.5). Fig. 12 shows summed positive DLA per operation and model. The dominant contributing component type varies across model families: Phi-3 is predominantly attention-driven, Gemma3-4B is MLP-dominant for most operations, Qwen3-1.7B shows mixed operation-dependent contributions, and Qwen2.5-3B exhibits relatively balanced effects. No single component type is uniformly responsible across all models and operations.

### D.6 Per-Direction Steering Results

Fig. 13 shows the underlying per-operation flip-to-target rates. The full per-comparison breakdown is provided in Tab. 10. A forward-inverse asymmetry is shown. In the forward regime, CAA directions achieve a non-tied win rate of 0.750 (12 wins from 16 non-tied comparisons across three steerable models), with SVD winning in 4 cases: Quantification for Phi-3 and Qwen2.5-3B, Co-hyponymy for Qwen2.5-3B, and Specification for Qwen3-1.7B. Three of these four SVD advantages are concentrated in lower-alignment operations (Quantification: $\cos = 0.459$; Co-hyponymy: $\cos = 0.514$), consistent with the principal variance direction pro-

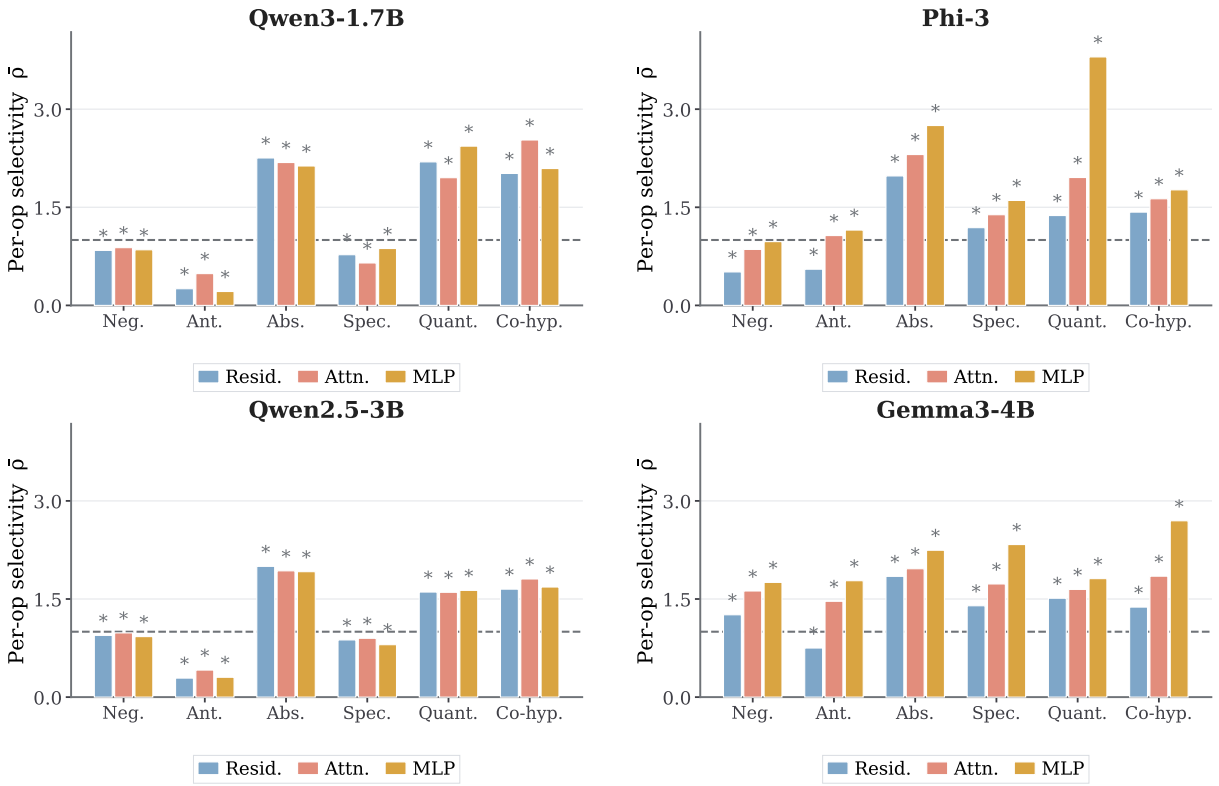


Figure 9: Per-operation selectivity ratio $\rho$ at the peak-selectivity layer for all four models. Abstraction exhibits the highest selectivity across models and components, followed by Quantification and Co-Hyponyms, while Negation and Antonymy show $\rho < 1$ in several configurations. The main paper reports the Phi-3 and Gemma3-4B subset.

Table 8: Best-layer linear classification accuracy for six-way operation classification (chance $= 1/6 = 0.167$). Mean and standard deviation are computed across five stratified cross-validation folds on held-out deltas projected into the per-operation subspaces.

| Model | Component | Layer | Accuracy | SD |
|---|---|---|---|---|
| Qwen-1.7B | `resid_post` | 9 | 0.944 | 0.015 |
| | `attn_out` | 26 | 0.934 | 0.017 |
| | `mlp_out` | 27 | 0.966 | 0.005 |
| Qwen2.5-3B | `resid_post` | 12 | 0.949 | 0.015 |
| | `attn_out` | 12 | 0.934 | 0.012 |
| | `mlp_out` | 1 | 0.959 | 0.015 |
| Phi-3 | `resid_post` | 31 | 0.989 | 0.005 |
| | `attn_out` | 12 | 0.956 | 0.007 |
| | `mlp_out` | 12 | 0.987 | 0.005 |
| Gemma-4B | `resid_post` | 7 | 0.999 | 0.001 |
| | `attn_out` | 25 | 0.998 | 0.003 |
| | `mlp_out` | 28 | 0.848 | 0.014 |

Table 9: Silhouette scores on t-SNE projections of activation deltas at the best-silhouette layer per model-component combination. All values are negative, confirming the absence of well-separated clusters in two-dimensional projection.

| Model | Component | Best layer | Silhouette |
|---|---|---|---|
| Qwen3-1.7B | `resid_post` | 25 | −0.176 |
| | `attn_out` | 3 | −0.168 |
| | `mlp_out` | 25 | −0.161 |
| Phi-3 | `resid_post` | 31 | −0.180 |
| | `attn_out` | 29 | −0.142 |
| | `mlp_out` | 30 | −0.186 |
| Qwen2.5-3B | `resid_post` | 31 | −0.230 |
| | `attn_out` | 32 | −0.038 |
| | `mlp_out` | 33 | −0.205 |
| Gemma3-4B | `resid_post` | 31 | −0.127 |
| | `attn_out` | 31 | −0.051 |
| | `mlp_out` | 31 | −0.134 |

viding a more concentrated causal signal when mean-shift and variance axes diverge substantially; the exception is Qwen3-1.7B Specification ($\cos = 0.810$, narrow margin: SVD 0.06, CAA 0.04). In the inverse regime, the non-tied win rate is 0.471 (8 CAA wins, 9 SVD wins from 17 non-tied comparisons), with a slight SVD advantage overall. The most striking pattern is a complete forward-inverse reversal for Qwen3-1.7B: CAA leads in the forward regime (win rate 0.667) but SVD wins all six operations in the inverse regime (win rate 0.000, $\bar{\Delta} = -0.337$), suggesting that the directions that best induce transformation effects are not the same

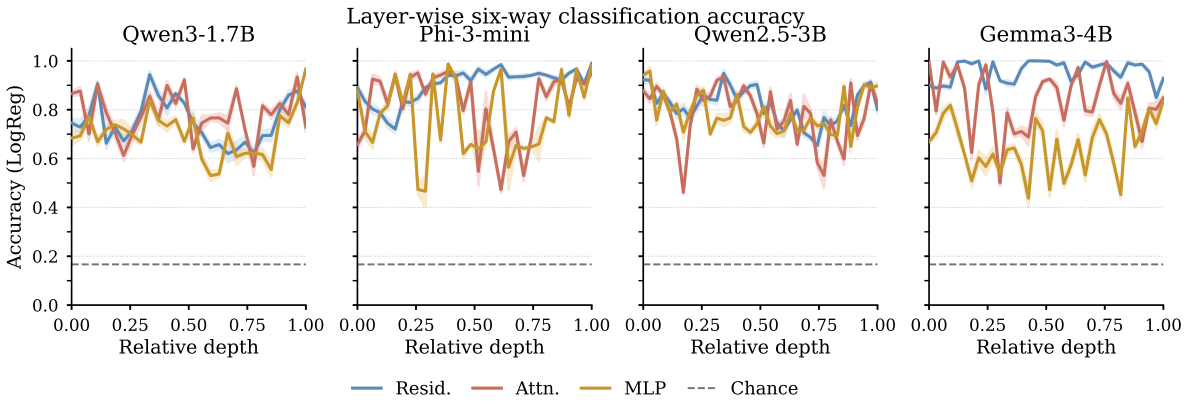


Figure 10: Layer-wise six-way classification accuracy (logistic regression) across relative depth for all four models and three components. Shaded bands show $\pm 1$ SD across five cross-validation folds. Dashed line marks chance ($1/6 \approx 0.167$). Accuracy rises sharply in early-to-mid layers and plateaus in upper layers, consistent with the selectivity profile in Fig. 3.

as those that best remove them. SVD wins in the inverse regime are concentrated in Qwen3-1.7B (six wins) and Qwen2.5-3B (three wins, including Quantification: SVD 1.00, CAA 0.00), while Phi-3 inverse steering is CAA-dominant across all operations.

### D.7 CAA vs SVD Direction Alignment

Fig. 14 shows the cosine similarity between CAA (mean-shift) and SVD (top singular vector) steering directions at five relative depth percentiles (0, 25, 50, 75, 100%) for each operation and model. Alignment varies substantially across models and operations, and evolves with depth. Phi-3 shows consistently high alignment across operations and layers, while Qwen2.5-3B shows strong operation-specific divergence (e.g. Co-hyponymy and Quantification remain near zero throughout). Early layers tend to show lower alignment, with convergence increasing toward the upper network for most configurations.

### D.8 Cross-Operation Interference: Per-Model and Inverse-Regime Results

Fig. 15 shows per-model cross-operation contamination heatmaps in the forward regime at $\alpha$=20, averaged over layers. The five high-interference pairs identified in the main text (contamination $> 0.07$) are model-dependent in their distribution. Qwen3-1.7B shows the most concentrated interference (6 of 30 off-diagonal pairs above 0.07), with the strongest effects involving Specification as source. Phi-3 shows broader interference (12 of 30 pairs above 0.07), concentrated among label-related pairs (Negation, Antonymy, Abstraction targeting Antonymy). Qwen2.5-3B exhibits the most diffuse pattern (20 of 30 pairs above 0.07), consistent with its weaker and more uniform per-operation selectivity profile in RQ1. Gemma3-4B shows near-zero contamination throughout the forward regime, reflecting the absence of reliable additive steering signal for this model. The inverse regime (Fig. 16) exhibits more variable patterns across models, with no consistent source $\rightarrow$ target structure at the aggregate level. We do not make strong structural claims about inverse-regime interference given this variability; the forward-regime pattern is the primary empirical basis for the RQ3 dissociation argument.

### D.9 Label-Controlled Operation Analysis

A possible confound in our setting is that operation identity is partially aligned with NLI label. Negation, Antonymy, and Quantification should produce contradiction examples, while Abstraction, Specification, and Co-hyponymy neutral examples. A direction that separates contradiction from neutral could therefore appear operation-selective even if it encodes only label-level structure. We address this directly with label-controlled analyses that test whether operation-specific structure remains when label identity is held fixed or explicitly removed.

**Method.** We use the same activation-difference and subspace notation as in §3.2. For operation $o$ at layer $l$, let $\mathbf{D}_o^{(l)} \in \mathbb{R}^{N \times d}$ denote the matrix of activation differences, and let $\mathbf{B}_o^{(l)} \in \mathbb{R}^{k \times d}$ be the row-stacked rank-$k$ basis obtained by mean-centred

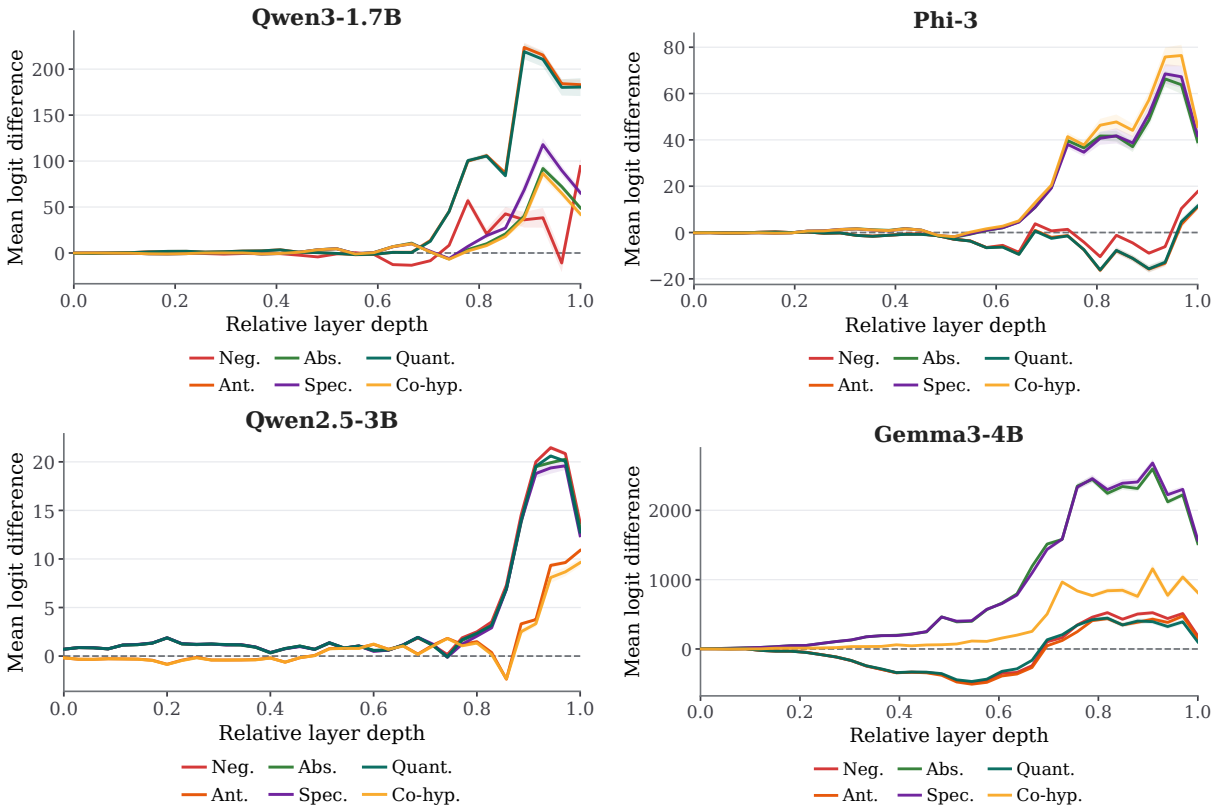


Figure 11: Logit lens trajectories across relative depth for all four models and all six operations. Positive values indicate that the decoded residual stream favours the target label induced by the transformation. Qwen3-1.7B, Phi-3, and Qwen2.5-3B show near-zero trajectories through early and middle layers, followed by sharp amplification in the upper network. Gemma3-4B exhibits a three-phase pattern: gradual early rise, mid-network suppression, and strong late amplification with substantially larger magnitudes.

SVD. Projection onto the operation subspace is defined as $eratornameproj_{\mathcal{S}_o^{(l)}}(\boldsymbol{\delta}) = \mathbf{B}_o^{(l)\top}\mathbf{B}_o^{(l)}\boldsymbol{\delta}$. Operations are grouped by label into contradiction and neutral sets. For each group, balanced samples are constructed by drawing equal numbers of activation differences from each operation, with replacement where necessary:

$$eratornameproj_{\mathcal{S}_o^{(l)}}(\boldsymbol{\delta}) = \mathbf{B}_o^{(l)\top}\mathbf{B}_o^{(l)}\boldsymbol{\delta}. \quad (13)$$

Operations are grouped by label into contradiction and neutral sets. For each group, balanced samples are constructed by drawing equal numbers of activation differences from each operation, with replacement where necessary.

We apply four controls. (i) *Within-label operation prediction*: operation identity is predicted while holding the NLI label fixed. (ii) *Label residualisation*: label-aligned components are removed by projecting activation differences orthogonally to a label subspace estimated via SVD. (iii) *Probe-depth comparison*: layer-wise peaks for label and operation prediction are compared. (iv) *Label-aware contamination*: cross-operation interference is compared between same-label and cross-label pairs.

**Within-label operation prediction.** Operation identity is recoverable from activation differences even when all examples share the same NLI label. Averaged across all model-component configurations, contradiction-only accuracy is $0.539$ and neutral-only accuracy is $0.655$, both well above the within-label chance rate of $1/3$ (Tab. 11). A classifier restricted to label-level information cannot exceed chance within a fixed-label group; above-chance performance therefore indicates the presence of operation-specific structure.

**Within-label selectivity.** Within-label selectivity ratios exceed $\rho = 1$ across all operations, models, and components (mean $\rho = 1.154$, minimum $\rho = 1.003$). This indicates that operation-specific structure persists within fixed-label groups, although with reduced separation consistent with the overlapping geometry observed in RQ1.

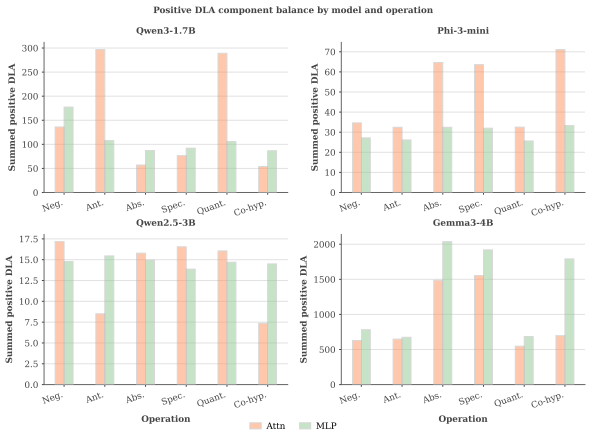


Figure 12: Summed positive direct logit attribution from attention and MLP components for each operation and model. Component contributions vary across architectures and operations, with no single component type uniformly dominant.

**Label residualisation.** Operation and label subspaces are strongly aligned (mean overlap 0.868), and label-subspace projections account for 87.8% of activation-difference variance. Despite this alignment, operation-specific structure remains after removing the label component. Mean selectivity increases from $\rho = 1.185$ to $\rho = 2.841$, and all residualised values remain above $\rho = 1$ (Tab. 12). This indicates that removing the shared label-aligned component increases the concentration of variance within operation-specific directions.

**Probe-depth comparison.** Peak depths for label prediction and operation prediction differ in some model-component configurations but coincide in others. For example, Phi-3 `attn` peaks at relative depth 0.968 for label prediction and 0.387 for operation prediction, while Qwen2.5-3B `resid` peaks at 0.971 and 0.400, respectively. Other configurations, including Gemma3-4B, show no separation. These differences are configuration-dependent and do not provide consistent evidence for depth-wise separation.

**Label-aware contamination.** If cross-operation interference were driven by label identity, same-label operation pairs would exhibit higher contamination than cross-label pairs. This pattern is not observed. Across all models, contamination levels are similar for same-label and cross-label pairs, with no statistically significant differences ($p \geq 0.721$) and small effect sizes (Tab. 13). Cross-operation interference is therefore not primarily explained by shared label identity.

Table 10: Full per-operation, per-model CAA vs SVD steering comparison. SVD and CAA columns report flip-to-target rates at the best $\alpha$ across all layers. Tie_zero: both directions produce zero. Tie_nonzero: both directions produce the same nonzero rate. Gemma3-4B rows are all Tie_zero and are omitted for space.

| Model | Operation | SVD | CAA | Winner |
|---|---|---|---|---|
| *Forward regime* | | | | |
| Phi-3 | Abstraction | 0.14 | 0.58 | CAA |
| | Antonymy | 0.74 | 0.78 | CAA |
| | Co-hyponymy | 0.12 | 0.16 | CAA |
| | Negation | 0.56 | 0.96 | CAA |
| | Quantification | 0.30 | 0.26 | SVD |
| | Specification | 0.08 | 0.46 | CAA |
| Qwen2.5-3B | Abstraction | 0.10 | 0.20 | CAA |
| | Antonymy | 0.00 | 1.00 | CAA |
| | Co-hyponymy | 1.00 | 0.98 | SVD |
| | Negation | 0.00 | 0.06 | CAA |
| | Quantification | 0.68 | 0.22 | SVD |
| | Specification | 0.96 | 0.96 | Tie (nonzero) |
| Qwen3-1.7B | Abstraction | 0.02 | 0.06 | CAA |
| | Antonymy | 0.06 | 0.20 | CAA |
| | Co-hyponymy | 0.14 | 0.14 | Tie (nonzero) |
| | Negation | 0.06 | 0.56 | CAA |
| | Quantification | 0.12 | 0.42 | CAA |
| | Specification | 0.06 | 0.04 | SVD |
| *Inverse regime* | | | | |
| Phi-3 | Abstraction | 0.50 | 1.00 | CAA |
| | Antonymy | 0.10 | 0.50 | CAA |
| | Co-hyponymy | 1.00 | 1.00 | Tie (nonzero) |
| | Negation | 0.00 | 0.28 | CAA |
| | Quantification | 0.00 | 0.06 | CAA |
| | Specification | 0.74 | 0.84 | CAA |
| Qwen2.5-3B | Abstraction | 1.00 | 0.92 | SVD |
| | Antonymy | 0.02 | 0.92 | CAA |
| | Co-hyponymy | 0.44 | 0.54 | CAA |
| | Negation | 1.00 | 0.98 | SVD |
| | Quantification | 1.00 | 0.00 | SVD |
| | Specification | 0.44 | 0.68 | CAA |
| Qwen3-1.7B | Abstraction | 0.76 | 0.22 | SVD |
| | Antonymy | 0.80 | 0.48 | SVD |
| | Co-hyponymy | 0.60 | 0.50 | SVD |
| | Negation | 0.62 | 0.30 | SVD |
| | Quantification | 1.00 | 0.28 | SVD |
| | Specification | 0.66 | 0.64 | SVD |
| **Forward non-tied win rate** | | | | CAA 0.750 / SVD 0.250 |
| **Inverse non-tied win rate** | | | | CAA 0.471 / SVD 0.529 |
| **Overall non-tied win rate (3 models)** | | | | CAA 0.606 / SVD 0.394 |

**Interpretation.** The four controls converge on the same conclusion. Operation-specific structure is not reducible to NLI label structure alone. Operation identity remains predictable within fixed-label groups, selectivity increases after removal of label-aligned components, and interference patterns do not track label identity. Although operation and label subspaces are strongly aligned, reflecting the design of the dataset, this alignment does not eliminate transformation-specific structure.

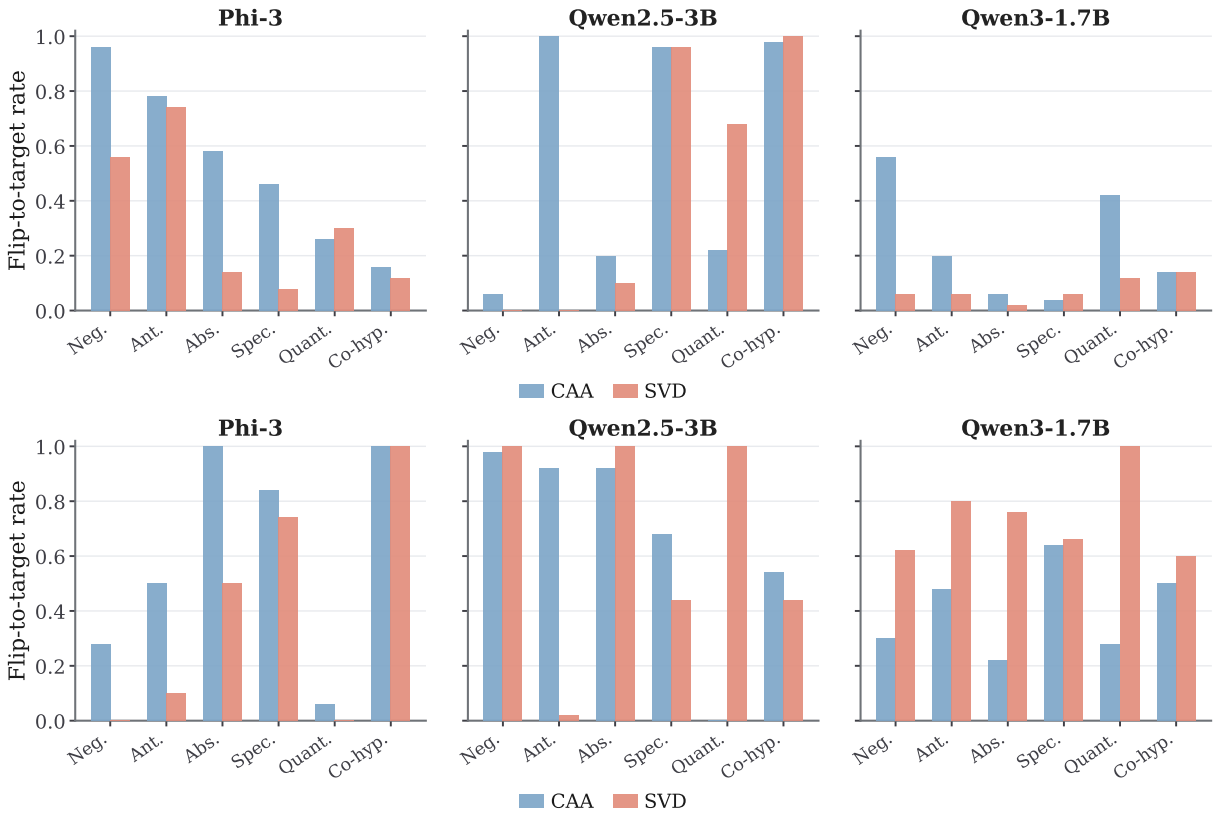


Figure 13: Per-operation flip-to-target rates for CAA mean-shift directions and SVD principal variance directions, shown separately for forward (induction; top) and inverse (removal; bottom) steering regimes. Each bar pair compares the two direction types for one operation in one model. Gemma3-4B is omitted because both direction types yield uniformly near-zero flip rates.

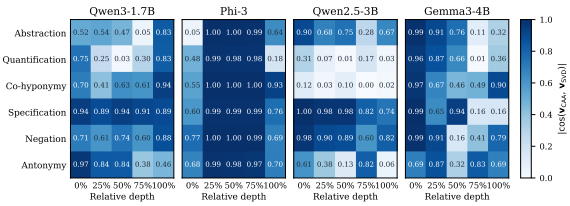


Figure 14: Layer-wise cosine similarity between CAA and SVD steering directions at relative depth percentiles 0/25/50/75/100%. Rows = operations (sorted by mean selectivity), columns = depth percentile, colour intensity = $|\cos(\mathbf{v}_{\mathrm{CAA}}, \mathbf{v}_{\mathrm{SVD}})|$.

Table 11: Within-label operation prediction accuracy. Each classifier distinguishes operations while holding the NLI label fixed. Chance is $1/3$ within each label group. Results are averaged over all twelve model-component configurations.

| Label group | Chance | Mean acc. | Range |
|---|---|---|---|
| Contradiction | 0.333 | 0.539 | 0.417-0.687 |
| Neutral | 0.333 | 0.655 | 0.523-0.780 |

Table 12: Effect of label-subspace residualisation on selectivity. Values are averaged across all operations, models, and components. Resid./raw reports the ratio $\rho_{\mathrm{resid}}/\rho_{\mathrm{raw}}$.

| Raw $\rho$ | Residualised $\rho$ | Resid./raw | Overlap | Label var. |
|---|---|---|---|---|
| 1.185 | 2.841 | 241.0% | 0.868 | 87.8% |

Table 13: Label-aware cross-operation contamination. Same-label pairs share the NLI label; cross-label pairs do not. No model shows a statistically reliable same-label amplification.

| Model | Same-label | Cross-label | $d$ | $p$ |
|---|---|---|---|---|
| Qwen3-1.7B | 0.036 | 0.054 | -0.150 | 0.721 |
| Phi-3 | 0.007 | 0.001 | 0.087 | 0.826 |
| Qwen2.5-3B | 0.100 | 0.100 | -0.004 | 0.991 |
| Gemma3-4B | -0.004 | -0.005 | 0.023 | 0.950 |

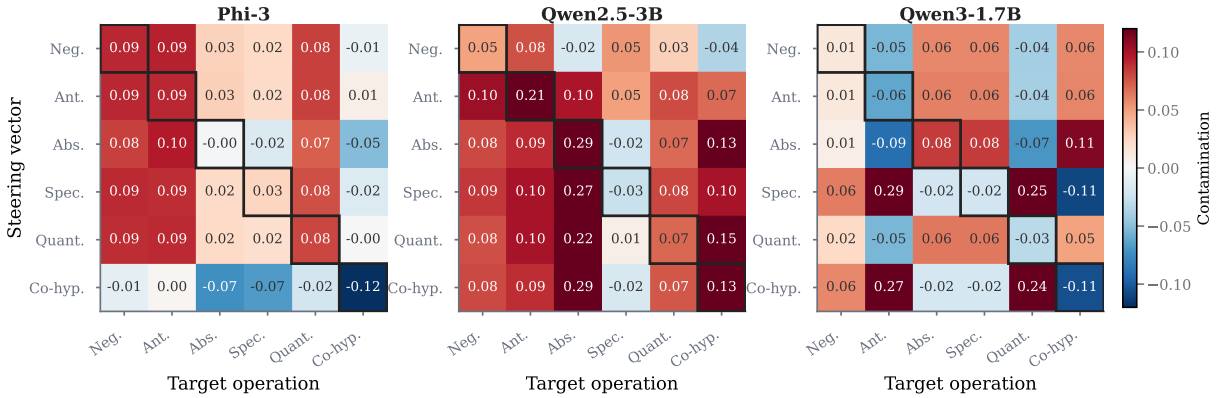


Figure 15: Per-model cross-operation steering contamination in the forward regime at $\alpha$=20, averaged over layers. Each panel corresponds to one model (three steerable models and Gemma3-4B). Rows index the source operation from which the SVD steering direction is derived; columns index the target operation whose held-out inputs are evaluated. Cell values give the excess flip-to-target rate relative to a matched random-direction baseline. The aggregate heatmap in the main paper (Fig. 4) reports the mean over all four models.

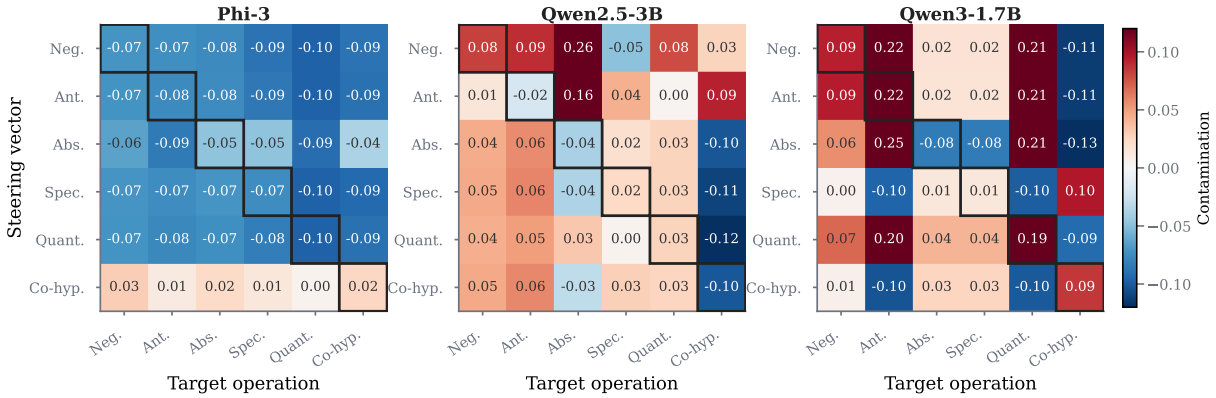


Figure 16: Per-model cross-operation steering contamination in the inverse regime at $\alpha$=20, averaged over layers. Patterns are more variable across models than in the forward regime, and no consistent source $\rightarrow$ target interference structure is observed at the aggregate level.