

Causal Transfer of Semantic Operators Across Transformer Language Models

Stanford CS224N Custom Project

Shivatmica Murgai
Department of Mathematics
Stanford University
smurgai@stanford.edu

Abstract

We study whether transformer language models represent relational knowledge as simple linear directions (“operators”) and whether these directions can be transferred across layers and model families to produce the intended behavioral effect. This is a stronger test than representational similarity alone: two internal spaces may appear geometrically aligned without being functionally interchangeable. Using a Wikidata-derived dataset of 4,350 facts and 73,720 templated prompts spanning 13 relations, we construct semantic atlases that summarize internal activations at selected layers and hook sites. We then perform three evaluations. First, we align entity representations across layers and models using PCA-space orthogonal Procrustes and measure pointwise retrieval with Recall@1, MRR, and mean diagonal cosine. Second, we test whether relation operators remain identifiable across the MLP boundary through nearest-neighbor retrieval. Third, we evaluate causal transport by injecting relation-swap directions and measuring the resulting change in token logit margin for the target object token. Within TinyLlama, pre-MLP representations are highly alignable across depth, whereas post-MLP updates are much less stable, suggesting that the residual stream provides a more consistent semantic coordinate system. Relation operators are identifiable far above chance in both TinyLlama and OLMo2, with the strongest performance in mid-depth layers. Within-model causal transport reliably outperforms random baselines, but remains weaker than native directions and degrades with greater layer distance. Additional relation-swap experiments show that the effect is not limited to a single GEO pair, although transfer strength varies substantially by relation pair. Cross-model alignment is moderate and sensitive to PCA rank, while cross-model causal transport remains much weaker than native.

1 Key Information to include

- Mentor: Qinan Yu
- External Collaborators (if you have any): None
- Sharing project: No
- Late Days (if you are using any): None

2 Introduction

A central goal of interpretability research is to understand not only whether large language models encode semantic structure internally, but whether those internal representations are functionally reusable across depth and across models. Prior work on representation similarity and probing has shown that transformer representations can exhibit strong geometric similarity across layers or architectures, but

such analyses are largely descriptive and do not by themselves establish functional equivalence [1, 2]. Mechanistic studies also suggest that semantic structure can differ across transformer components and layer depth [3, 4].

Most prior work studies concept-level directions [5, 6], but many forms of language understanding depend on structured relations. If relation-specific directions can be transported across layers or models and still induce the intended behavioral change, that would indicate a stronger form of internal consistency than representational similarity alone.

For intuition, consider the prompt template “France is on the continent of ____.” A successful `continent_of` $\rightarrow$ `official_language_of` transport should decrease the model’s preference for the source-relation answer (Europe) and increase its preference for the target-relation answer (French). In this project, I test whether transformer language models encode relations as reusable operator-like directions. I construct semantic atlases from a Wikidata-derived dataset [7] by aggregating last-token activations over relation-conditioned prompts at selected layers and hook sites. From these atlases, I derive one mean-centered, unit-normalized direction per entity and per relation. Operationally, transport is successful when injecting a transported swap direction increases the logit margin of the target relation’s object token over the source relation’s object token.

Concretely, the project asks three related questions. First, do entity representations align across layers and across models closely enough to define a shared semantic geometry? Second, are relation directions identifiable across the MLP boundary, suggesting that they persist as recognizable operators from pre- to post-MLP states? Third, when these directions are transported and injected, do they causally shift model behavior as intended? These questions will help differentiate between geometric similarity and functional portability.

The main results support a clear mechanistic picture. In TinyLlama, entity geometry is highly stable at `mlp_input` but substantially less stable at `mlp_output`. Relation operators are strongly identifiable across the MLP boundary in both TinyLlama and OLMo2, with the strongest performance in intermediate layers. Within-model causal transport succeeds reliably and shows a distance effect, where retention falls from about 0.94 for near-layer transport to about 0.878 for farther-layer transport, even when geometric alignment remains strong. Additional relation-swap experiments show that the effect is not confined to a single GEO relation pair, although transfer strength varies substantially across relation pairs. Cross-model alignment is moderate and improves with PCA rank, but cross-model causal transport remains much weaker than native.

3 Related Work

This project sits at the intersection of representation similarity, activation steering, and mechanistic interpretability. Work on representation similarity studies when hidden states from different layers or models occupy comparable geometric structures, using tools such as correlation-based comparison or explicit alignment in shared subspaces [1, 2]. These methods are useful for describing similarity, but they do not by themselves show that an aligned direction in one space can be used as a functional intervention in another.

The closest prior work is Huang et al., who study cross-model transferability of concept directions by asking whether a direction learned in one model can still steer another model’s behavior as intended [6]. My project adopts that behavioral notion of transfer, but extends it from single concepts to relational operators. It also relates to activation-steering work showing that linear directions can causally influence model outputs [5].

Finally, this project is motivated by mechanistic work suggesting that semantic structure is not uniformly represented across layers and components [3, 4]. That perspective motivates the focus here on depth, hook location, and the MLP boundary. Relative to prior work, the present study asks when aligned relational directions are not just recognizable, but reusable enough to transport causal behavior across depth or across model families.

4 Approach

My goal is to test whether relations are represented in transformer language models as reusable linear operators, and whether those operators can be aligned, identified, and causally transported across layers and models.

4.1 Semantic atlases

Intuitively, a semantic atlas is a lookup table of one vector per entity and one vector per relation at a fixed model, layer, and hook site. I build it by averaging last-token activations over all prompts involving that entity or relation, subtracting a global mean, and normalizing the resulting vectors.

I record internal activations at two hook sites: `mlp_input`, the residual stream immediately before the MLP block, and `mlp_output`, the MLP update vector before it is added back into the residual stream. At a fixed model, layer, and hook site, the atlas therefore provides one mean-centered, unit-normalized direction per entity and per relation.

The dataset consists of fact triples $\langle e, r, t \rangle$, where e is the subject entity, r is the relation, and t is the target object entity, rendered into relation-specific natural-language prompts. For example, a fact such as (France, `continent_of`, Europe) may be rendered with a template such as "France is on the continent of." For each prompt instantiation $p(e, r, t)$, I run a forward pass and extract the activation at the final non-padding token, denoted $x_{\ell,h}(e, r, t) \in \mathbb{R}^d$ for layer ℓ and hook h .

Let P denote the set of all instantiated prompts used to build the atlas. For each layer ℓ and hook h , I compute a global mean over all instantiated prompts,

$$\mu_{\ell,h} = \mathbb{E}_{(e,r,t) \in P} [x_{\ell,h}(e, r, t)].$$

I then compute per-entity and per-relation means to get

$$\bar{x}_{\ell,h}(e) = \mathbb{E}_{(r,t): (e,r,t) \in P} [x_{\ell,h}(e, r, t)], \quad \bar{x}_{\ell,h}(r) = \mathbb{E}_{(e,t): (e,r,t) \in P} [x_{\ell,h}(e, r, t)].$$

Entity and relation directions are formed by mean-centering, where

$$d_{e,\ell,h} = \bar{x}_{\ell,h}(e) - \mu_{\ell,h}, \quad d_{r,\ell,h} = \bar{x}_{\ell,h}(r) - \mu_{\ell,h}.$$

Finally, I apply row-wise ℓ_2 normalization and this gives

$$\hat{d}_{e,\ell,h} = \frac{d_{e,\ell,h}}{\|d_{e,\ell,h}\|_2}, \quad \hat{d}_{r,\ell,h} = \frac{d_{r,\ell,h}}{\|d_{r,\ell,h}\|_2}.$$

4.2 Alignment and retrieval

To compare semantic structure across layers or across models, I align entity direction matrices using PCA followed by orthogonal Procrustes. PCA reduces each set of entity directions to a lower-dimensional coordinate system that captures the main variation, making the two spaces easier to compare. Given two entity direction matrices $D_A \in \mathbb{R}^{|E| \times d_A}$ and $D_B \in \mathbb{R}^{|E| \times d_B}$, where E is the set of matched entities used for alignment, I fit PCA separately and project each into a rank- k subspace, obtaining $S_A = \text{PCA}_k(D_A)$ and $S_B = \text{PCA}_k(D_B)$.

Orthogonal Procrustes then finds the best rotation between these reduced spaces so that matched entities line up as closely as possible, where $Q^* = \arg \min_{Q \in O(k)} \|S_A - S_B Q\|_F$.

After alignment, I test whether each source entity’s aligned vector retrieves its correct counterpart from the full pool of target entity directions under cosine nearest-neighbor search. The primary retrieval metrics are Recall@1 and mean reciprocal rank (MRR). I also report mean diagonal cosine after Procrustes, defined as the average row-wise cosine similarity between aligned source entities and their matched targets.

4.3 Operator identifiability across hooks

For operator identification, each source relation direction is matched against the full set of 13 target relation directions at the paired hook using cosine similarity. At a fixed layer, I retrieve each relation direction in `mlp_input` from the set of relation directions in `mlp_output` using cosine similarity. With 13 relations in the dataset, chance Recall@1 is $1/13 \approx 7.69\%$.

4.4 Causal operator transport

Let $S(r)$ denote the PCA-space representation of relation r in the aligned source or target semantic space. For a swap $r_s \rightarrow r_t$, I define a swap direction in PCA space as $\Delta = S(r_t) - S(r_s)$. Given an intervention direction v , I inject αv at the last-token position, where α is a scalar controlling intervention strength.

I compare three intervention directions:

1. **Random**: a random unit vector baseline in the target hidden space.
2. **Transfer**: the swap direction learned in the source model, mapped into the target model’s PCA-aligned coordinate system using the learned Procrustes transform. After transport in PCA space, I map the direction back through the target PCA basis to obtain a vector in the target model’s original hidden space, which is the vector used for intervention.
3. **Native**: the swap direction computed directly from the target atlas.

I inject the chosen direction at the last-token position and measure its effect on the model’s preference for the target object token:

$$m(p) = z_\tau(p) - z_\sigma(p), \quad \Delta m = m_{\text{after}} - m_{\text{before}},$$

where τ is the single-token object corresponding to the target relation completion in the one-token evaluation subset, and σ is the single-token object corresponding to the source relation completion. I report mean Δm , $\text{Win}(T > R)$, and

$$\text{Retention} = \frac{\mathbb{E}[\Delta m_{\text{Transfer}}]}{\mathbb{E}[\Delta m_{\text{Native}}]}.$$

$\text{Win}(T > R)$ is the fraction of evaluation examples for which the transferred direction produces a larger logit-margin gain than the random baseline:

$$\text{Win}(T > R) = \frac{1}{N} \sum_{i=1}^N \mathbf{1}[\Delta m_{\text{Transfer}}^{(i)} > \Delta m_{\text{Random}}^{(i)}].$$

An aligned direction is only meaningful if it improves the target margin more reliably than an equally sized arbitrary intervention.

5 Experiments

I evaluate the semantic-atlas pipeline on a Wikidata-derived factual dataset using retrieval, identifiability, and causal-transport experiments.

5.1 Data

I construct a Wikidata-derived dataset from a subset of factual triples and relation-specific prompt templates [7]. The dataset contains entities, relations, prompt templates, a fact table, dataset metadata, and type-specific entity groups that ensure that each relation/template pair is matched with appropriate entities. The full fact table contains 4,350 rows of the form $(e, r \rightarrow t)$, and atlas extraction produces 73,720 instantiated prompts spanning 13 relations.

Prompts are instantiated only from observed (entity, relation) pairs in the dataset, with each relation using its own set of templates applied to those observed pairs rather than to unseen entity-relation combinations. For causal transport, I additionally use a one-token subset of facts since the logit-margin metric compares candidate object tokens directly. Causal evaluation is restricted to one-token objects so that the intervention effect can be measured as a logit-margin difference between the source and target answer tokens. The main within-model causal experiment uses the swap `continent_of` $\rightarrow$ `official_language_of` on 82 overlapping one-token entities.

5.2 Evaluation method

For entity alignment, I report Recall@1, MRR, and mean diagonal cosine after Procrustes. For operator identifiability, I report relation Recall@1 across the MLP boundary. For causal transport, I report mean logit-margin change Δm , $\text{Win}(T > R)$, and Retention. Together, these metrics distinguish directions that are merely aligned from directions that are behaviorally effective.

5.3 Experimental details

I evaluate two pretrained transformer language models: TinyLlama/TinyLlama-1.1B-intermediate-step-1431k-3T [8] and allenai/OLMo-2-1124-7B [9]. I chose a small set of mid-to-late layers that are spaced closely enough to test within-model depth transport, and paired them across TinyLlama and OLMo2 as approximate matched-depth anchors for cross-model comparison. For TinyLlama, I extract atlases at layers {16, 18, 19, 21}; for OLMo2, I extract atlases at layers {24, 26, 28, 31}. Across models, these layer pairs should be understood as approximate matched-depth anchors based on relative depth within each stack, not as a claim of exact functional correspondence. At each layer, I record activations at two hook sites, `mlp_input` and `mlp_output`. Unless otherwise stated, PCA rank is $k = 64$; for PCA-rank sensitivity, I also evaluate $k \in \{32, 128\}$.

All experiments are purely analytical: no fine-tuning is performed, and model parameters remain fixed throughout. The overall pipeline consists of forward-pass activation extraction, semantic-atlas construction, PCA/Procrustes-based alignment, nearest-neighbor retrieval, and causal intervention. I implemented all the code for atlas extraction, alignment, retrieval, and intervention.

For entity retrieval, each aligned source entity is matched against the full set of target entity directions.

For within-model causal transport, the main setup injects at TinyLlama layer 21 using a source direction from either layer 19 or layer 16, allowing comparison between near-layer and far-layer transport. The main relation swap is `continent_of` $\rightarrow$ `official_language_of`, and I sweep $\alpha \in \{0.5, 1, 2\}$. I also test additional swaps including `capital_of` $\rightarrow$ `currency_of` and `capital_of` $\rightarrow$ `official_language_of`. For cross-model causal transport, I evaluate OLMo2 L26 $\rightarrow$ TinyLlama L18.

In the operator-identification task, each relation is represented by a single linear atlas direction at a fixed layer/hook, and we test whether its `mlp_input` direction retrieves the matching `mlp_output` relation from the pool of all 13 target relation directions by cosine similarity.

For cross-model transfer, I first map both models’ entity directions into a shared PCA subspace, fit an orthogonal Procrustes transform using matched entities, and then apply that transform to transport source-model relation-swap directions into the target model’s coordinate system before intervention.

5.4 Results

The results support a consistent picture across retrieval and intervention experiments: pre-MLP residual geometry is highly stable, post-MLP update space is much less stable, relation operators are identifiable across the MLP boundary, and causal transport works within model but attenuates with distance and across model families.

5.4.1 Within-model entity alignment in TinyLlama

Table 1 shows that TinyLlama `mlp_input` supports a highly stable entity geometry across depth: Recall@1 is 98–100% across nearby layers and remains 91.67% even for the larger 16 $\rightarrow$ 21 gap. In contrast, `mlp_output` is substantially less stable, with Recall@1 ranging from 40.42% to 72.81%. Mean diagonal cosine shows the same pattern (Table 2). These results suggest that the pre-MLP residual stream serves as a relatively stable semantic coordinate system, whereas the post-MLP update space appears less structurally consistent and more specific to individual layers.

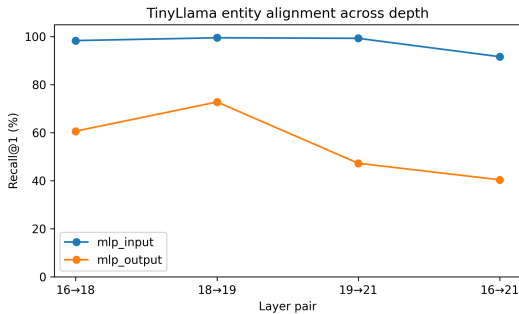


Figure 1: TinyLlama entity alignment across depth

Table 1: TinyLlama within-model entity alignment after PCA- k Procrustes ($k = 64$).

Hook	Pair	ΔL	Recall@1	MRR
mlp_input	16→18	2	98.38%	0.9903
mlp_input	18→19	1	99.55%	0.9975
mlp_input	19→21	2	99.32%	0.9959
mlp_input	16→21	5	91.67%	0.9442
mlp_output	16→18	2	60.67%	0.7022
mlp_output	18→19	1	72.81%	0.8052
mlp_output	19→21	2	47.29%	0.6051
mlp_output	16→21	5	40.42%	0.5360

Table 2: Mean diagonal cosine after Procrustes across TinyLlama layers ($k = 64$).

Pair (tgt/src)	MeanCos (mlp_input)	MeanCos (mlp_output)
18/16	0.6478	-0.1005
19/18	0.7233	0.3993
21/19	0.6627	0.4364
21/16	0.4233	0.4086

5.4.2 Cross-model entity alignment and PCA-rank sensitivity

Cross-model entity alignment between TinyLlama and OLMo2 exists but is substantially weaker than within-model alignment: at matched relative depths and $k = 64$, Recall@1 is roughly 45 – 53%. Cross-model correspondence is also sensitive to PCA rank. For instance, on the Tiny18 → OLMo26 pair, increasing k from 32 to 64 to 128 improves Recall@1 from 40.72% to 52.79% to 58.97%, and MRR from 0.5229 to 0.6292 to 0.6825.

5.4.3 Operator identifiability across the MLP boundary

Relation operators are strongly identifiable across the MLP boundary in both models (Table 3). With 13 relations, chance performance is only $1/13 \approx 7.69\%$, but observed Recall@1 is far higher. The strongest identifiability appears in mid-depth layers, suggesting that relation structure is represented most cleanly at intermediate depth.

5.4.4 Within-model causal transport and layer-distance effect

The main within-model causal experiment uses the swap `continent_of` → `official_language_of` on 82 overlapping one-token entities, with interventions at TinyLlama layer 21 and $k = 64$. Near-layer transport from source layer 19 succeeds strongly: at $\alpha = 1.0$, Random is -0.0131 , Transfer is 0.6492 , and Native is 0.6909 , with $\text{Win}(T>R) = 100\%$ and Retention = 0.940 . Repeating the same intervention with the farther source layer 16 instead of 19 still succeeds strongly, but retention drops to 0.878 . This means that even when entity geometry remains alignable, causal operator transport is not fully interchangeable across depth (Table 4).

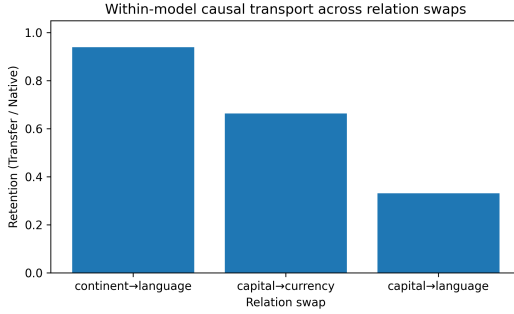


Figure 2: Retention at $\alpha = 1$ for successful within-model relation swaps

5.4.5 Additional relation-swap robustness

To test whether the causal effect is limited to a single GEO swap, I evaluated additional relation pairs under the same TinyLlama L19 → L21, `mlp_output`, $k = 64$ setup (Table 5). The strongest additional result is `capital_of` → `currency_of`: at $\alpha = 1$, this swap uses 283 entities and yields Random $\Delta m = -0.1110$, Transfer

Table 3: Operator identifiability across hooks (Relation Recall@1; chance = 7.69%).

TinyLlama		OLMo2	
Layer	Relation Recall@1	Layer	Relation Recall@1
16	76.92%	24	100.00%
18	100.00%	26	100.00%
19	100.00%	28	61.54%
21	61.54%	31	84.62%

Table 4: Within-model causal transport on 82 entities (TinyLlama; injection at `mlp_output`; $k = 64$).

Setting	α	Random	Transfer	Native	Win(T>R)	Retention
L19→L21	0.5	-0.0080	0.3217	0.3424	100%	0.940
L19→L21	1.0	-0.0131	0.6492	0.6909	100%	0.940
L19→L21	2.0	-0.0146	1.3213	1.4062	100%	0.940
L16→L21	0.5	-0.0080	0.3013	0.3431	100%	0.878
L16→L21	1.0	-0.0131	0.6081	0.6924	100%	0.878
L16→L21	2.0	-0.0146	1.2376	1.4092	100%	0.878

$\Delta m = 0.1319$, and Native $\Delta m = 0.1987$, with Win(T>R) = 100.0% and Retention = 0.664. A second additional swap, `capital_of` → `official_language_of`, is weaker but still above baseline, with Win(T>R) = 76.4% and Retention = 0.331.

Not all swaps succeed under the same protocol. For example, `currency_of` → `official_language_of` is weak, with Transfer below Random on average and Win(T>R) = 35.4%, and `continent_of` → `currency_of` is even weaker, with Win(T>R) = 16.7%. These weak cases show that the transport phenomenon is not uniformly robust across all semantic transitions.

5.4.6 Cross-model causal transport

Finally, I test cross-model causal transport from OLMo2 to TinyLlama. Using the swap `continent_of` → `official_language_of` and transporting a direction from OLMo2 layer 26 to TinyLlama layer 18, transfer is above random but much weaker than native. At $\alpha = 1.0$, Random is 0.0313, Transfer is 0.0902, and Native is 0.4672, with Win(T>R) = 58.5% and Retention = 0.193. These results show that cross-model geometric alignment is not sufficient for strong cross-model functional portability.

6 Analysis

Beyond the aggregate results, the experiments reveal a consistent mechanistic distinction between semantic geometry and functional transport.

6.1 Pre-MLP residual geometry vs post-MLP update space

The results display a clear contrast between `mlp_input` and `mlp_output`. At `mlp_input`, entity retrieval across nearby layers in TinyLlama is essentially near-bijective, with Recall@1 above 98% for moderate gaps and still above 91% even for the larger 16 → 21 jump. At `mlp_output`, the same comparisons drop dramatically, in some cases to around 40 – 47%. This suggests that the pre-MLP residual stream behaves more like a stable semantic coordinate system, while the post-MLP update space is less coordinate-like and more layer-specific.

6.2 Geometry is not the same as function

Strong geometric alignment does not imply full functional interchangeability. Within TinyLlama, entity geometry remains highly alignable even across larger layer gaps, but causal retention still falls when transporting from a farther source layer.

Table 5: Additional relation-swap robustness at $\alpha = 1$ (TinyLlama L19→L21, mlp_output, $k = 64$).

Swap	N	Random Δm	Transfer Δm	Native Δm	Win(T>R)	Retention
continent_of → official_language_of	82	-0.0131	0.6492	0.6909	100.0%	0.940
capital_of → currency_of	283	-0.1110	0.1319	0.1987	100.0%	0.664
capital_of → official_language_of	423	-0.0300	0.0267	0.0806	76.4%	0.331

In the main continent_of → official_language_of experiment, retention drops from about 0.940 for 19 → 21 transport to about 0.878 for 16 → 21 transport. A direction can therefore remain geometrically meaningful without being fully interchangeable as an operator at another layer depth.

6.3 Successes and failures across relation pairs

The additional relation-swap experiments show that causal transport is not an artifact of one hand-picked GEO pair, but they also reveal that success varies substantially across relation pairs. capital_of → currency_of is a strong additional success case; capital_of → official_language_of is weaker but still above baseline. In additional exploratory swap runs, some other pairs such as currency_of → official_language_of and continent_of → currency_of were much weaker, suggesting that transport robustness varies sharply by relation pair. This variation suggests that some relation transitions are much more reusable than others, and that strong operator identifiability does not automatically imply equally strong causal transport.

More broadly, these results suggest that operator identifiability is a necessary but not sufficient condition for strong causal portability, where a relation direction can remain recognizable across representations without yielding equally strong transferred behavior under intervention.

6.4 Prompt-level intuition

For a concrete example, consider a prompt built from the fact (France, continent_of, Europe), rendered as “France is on the continent of ____.” Under a successful continent_of → official_language_of transport intervention, the desired behavior is not merely that the source and target spaces look aligned, but that the intervention shifts probability mass away from the source-relation completion and toward the target-relation completion. This example captures the key distinction that motivates the paper: geometric alignment is about whether directions can be matched, while causal transport is about whether the matched direction actually changes the model’s behavior in the intended semantic direction.

7 Conclusion

This project studied whether transformer language models encode relations as reusable operator-like directions and whether those directions can be aligned, identified, and causally transported across layers and model families. Using the semantic atlases, I found that entity geometry is highly stable in the pre-MLP residual stream but less stable in the post-MLP update space. Relation operators remain strongly identifiable across the MLP boundary in both TinyLlama and OLMo2, especially in mid-depth layers. Within-model causal transport succeeds reliably, beats random baselines, and remains weaker than native, while layer-distance experiments show that functional transport degrades even when geometric alignment remains strong. Additional relation-swap experiments show that the effect is not limited to a single GEO pair, though robustness varies across relation pairs. Cross-model alignment is moderate and improves with rank, but causal transport remains much weaker than native.

Overall, these results support a clear distinction between semantic geometry and semantic function. The residual stream appears to provide a stable semantic coordinate system, but functional portability is more limited than geometric similarity suggests. This study is intentionally scoped to a controlled setting: causal evaluation is restricted to one-token objects, experiments focus on a selected set of layers and hook sites, and the analysis emphasizes the transfer pipeline. Future work should extend the analysis to more relations and templates, test additional intervention sites such as attention-side hooks, relax the single-token evaluation constraint, and analyze why some relation pairs remain identifiable yet fail to transfer strongly.

References

- [1] Olha Sirikova and Alvin Chan. The triangle of similarity: A multi-faceted framework for comparing neural network representations. *arXiv preprint*, 2026.
- [2] Anirudh Achara et al. Multi-way representation alignment. *arXiv preprint*, 2026.
- [3] Oscar Skean, Md Rifat Arefin, Dan Zhao, Niket Patel, Jalal Naghiyev, Yann LeCun, and Ravid Shwartz-Ziv. Layer by layer: Uncovering hidden representations in language models. *arXiv preprint*, 2025.
- [4] Shawn Im, Changdae Oh, Zhen Fang, and Sharon Li. How do transformers learn to associate tokens: Gradient leading terms bring mechanistic interpretability. *ICLR*, 2026.
- [5] Samuel Soo, Wesley Teng, and Chandrasekaran Balaganesh. Interpretable steering of large language models with feature guided activation additions. *arXiv preprint*, 2025.
- [6] Rui Huang et al. Cross-model transferability among large language models on the platonic representations of concepts. *arXiv preprint*, 2025.
- [7] Wikidata: A free collaborative knowledge base. <https://www.wikidata.org/>. Accessed 2026-02-20.
- [8] Tinyllama-1.1b-intermediate-step-1431k-3t. <https://huggingface.co/TinyLlama/TinyLlama-1.1B-intermediate-step-1431k-3T>. Model card.
- [9] Olmo-2-1124-7b. <https://huggingface.co/allenai/OLMo-2-1124-7B>. Model card.