Layer-wise Embedding-Output Dynamics Across LLM Families: Evidence for Phase-Structured Decision Commitment

Davide D'Elia[*]
IU International University of Applied Sciences
davide.delia@iu-study.org

Abstract

We investigate the layer-wise relationship between embedding geometry and output preferences across four diverse LLM families: Pythia-6.9B, Llama-3.1-8B, Apertus-8B (multilingual), and Gemma-2B (SFT). Using a pair-level centroid-asymmetry metric computed over 230 statement pairs with 10,000 bootstrap resamples, we find that embedding-output dynamics are **phase-structured**: the correlation between embedding clustering and output preference changes sign systematically across network depth.

In three out of four models, we observe significant late-layer inversion ($r = -0.17$ to -0.41, 95% CI excluding zero, $p < 0.001$), while early layers **in base and multilingual models** show positive correlations ($r = +0.27$ to $+0.49$). The transition point varies by architecture: Layer 28-32 for Pythia, Layer 4-8 for Llama, and Layer 12 for Apertus. Notably, Apertus-8B shows maximal inversion at Layer 28 ($r = -0.41$) rather than the final layer, suggesting decision commitment may not be uniformly localized at network output.

Gemma-2B shows no significant late-layer inversion, representing a **boundary condition** potentially related to scale or training method. Llama-3.1-8B-Instruct with chat templates shows uniformly negative correlations ($r = -0.47$ to -0.60), suggesting instruction tuning **is associated with** a distinct processing regime activated by deployment-format inputs. Our findings demonstrate that late-layer inversion is architecturally robust in larger base models, while the depth of decision commitment and training methodology effects require further causal investigation.

1 Introduction

Large language models (LLMs) process information through dozens of transformer layers, progressively transforming input embeddings into output distributions. Understanding *where* and *how* decisions crystallize within this processing pipeline is crucial for mechanistic interpretability and AI safety research.

Prior work has examined layer-wise representations through probing classifiers (Belinkov, 2022), attention pattern analysis, and residual stream decomposition. However, a fundamental question remains underexplored: *How does the relationship between embedding geometry and output preference evolve across network depth?*

We address this question through systematic layer-wise analysis across four architecturally diverse LLM families. Our key finding is that embedding-output dynamics are **phase-structured**—the correlation between embedding clustering and output preference exhibits systematic sign changes across depth, rather than remaining constant or monotonically varying.

1.1 Research Questions

1. Is the embedding-output relationship consistent across layers?
2. Does this relationship vary across model families?
3. How do training methods (base vs. SFT vs. RLHF) affect this relationship?
4. Is there a universal "decision point" in LLM processing?

1.2 Contributions

Our main contributions are:

1. **Empirical Finding:** We demonstrate that embedding-output relationships exhibit phase-structured dynamics in base and multilingual

[*] This work was conducted independently in a private capacity and does not represent the views of IU International University of Applied Sciences.

models, with early layers showing positive correlation ($r = +0.27$ to $+0.49$) and late layers showing inversion ($r = -0.17$ to -0.41) in 3/4 tested model families.

2. **Cross-Model Validation:** To our knowledge, we provide the first systematic comparison of layer-wise embedding-output dynamics across four architecturally diverse models, establishing that late-layer inversion is robust but architecture-dependent in depth.

3. **Boundary Case Identification:** We identify conditions where phase structure does not reliably emerge: small-scale SFT models (Gemma-2B, 2B parameters) lack significant late-layer inversion, representing a boundary condition that may relate to scale, layer count, or supervision method.

4. **Methodological Contribution:** We introduce a rigorous pair-level analysis framework ($n = 230$, bootstrap $n = 10,000$) that resolves signals masked by category-level aggregation due to Simpson's Paradox effects.

2 Related Work

2.1 Layer-wise Analysis in Transformers

Probing classifiers have revealed that different layers encode different types of information, with syntactic information typically peaking in middle layers and semantic information in later layers (Belinkov, 2022; Tenney et al., 2019). Our work complements this by examining not *what* layers encode, but how encoding *relates to output preferences*.

2.2 Embedding Geometry

Recent work on the "Platonic Representation Hypothesis" (Huh et al., 2024) suggests that models trained on different modalities converge toward similar geometric structures. The geometric memory hypothesis (Noroozizadeh et al., 2025) demonstrates that models store facts through synthesized embeddings encoding global relationships. Our work examines how this geometry relates to output preferences *across layers*.

2.3 Representational Preferences

Prior work introduced *Uniformity Asymmetry* as an exploratory metric for detecting representational preferences in LLM embeddings (D'Elia, 2025). That work outlined three directions for future research: (1) output correlation, (2) layer-wise analysis, and (3) downstream validation. The present study directly addresses the first two directions, extending the metric to pair-level analysis and systematically examining layer-wise dynamics.

2.4 Decision Making in LLMs

Concurrent work by Ganesh et al. (2025) demonstrates through KV-cache manipulation that "high-level structural plans are encoded early in the generation process" while "local discourse structure is maintained by final layers." Our embedding geometry analysis provides converging evidence for this early-planning, late-execution distinction through an independent methodology.

2.5 Instruction Tuning and Representation Engineering

RLHF and instruction tuning modify model behavior (Ouyang et al., 2022), but the representational changes remain poorly understood. Recent work on representation engineering (Zou et al., 2023) shows that activation steering can modify model outputs. Our template comparison provides evidence that instruction tuning is associated with distinct processing regimes conditional on input formatting.

2.6 Layer-wise Processing in Safety Contexts

Understanding where decisions crystallize has implications for AI safety. If early layers encode "honest" representations that are transformed in later processing, interventions targeting wrong layers may be ineffective (Elhage et al., 2021; Geva et al., 2023).

3 Method

3.1 Dataset

We use a dataset of 230 statement pairs across six semantic categories. Each pair consists of Statement A (framed as more uniform/general) and Statement B (framed as more specific/hedged). **Important:** This framing is a controlled experimental design, not a normative claim about truth or quality. Categories include ground-truth numeric facts, taste preferences, proverbs, and contested claims.

3.2 Models

We analyze four models selected for architectural diversity:

Model	Layers	Params	Type
Pythia-6.9B (Biderman et al., 2023)	32	6.9B	Base
Llama-3.1-8B (Touvron et al., 2023)	32	8B	Base
Apertus-8B	32	8B	Multiling.
Gemma-2B (Team et al., 2024)	18	2B	SFT

Table 1: Models analyzed. Selection criteria: architectural diversity, open weights, and boundary case inclusion.

3.3 Metrics

Output Preference. For each statement pair, we compute output preference as the difference in negative log-likelihood:

$$\text{pref}(A, B) = \text{NLL}(B) - \text{NLL}(A) \qquad (1)$$

Positive values indicate the model assigns lower perplexity (higher preference) to Statement A.

Centroid Asymmetry (Pair-Level). Following D'Elia (2025), we compute centroid-based asymmetry measures. For each layer l, we extract mean-pooled hidden states for all statements:

$$\mathbf{e}_i^{(l)} = \frac{1}{|T_{\text{valid}}|} \sum_{t \in T_{\text{valid}}} \mathbf{h}_{i,t}^{(l)} \qquad (2)$$

Pooling specification: We mean-pool across non-special tokens, excluding BOS and system tokens where applicable. This pooling strategy is held constant across all layers and models.

We compute class centroids $\mathbf{c}_A^{(l)}$ and $\mathbf{c}_B^{(l)}$ over all pairs, then measure asymmetry:

$$\text{asym}_i^{(l)} = \cos(\mathbf{e}_{A,i}^{(l)}, \mathbf{c}_A^{(l)}) - \cos(\mathbf{e}_{B,i}^{(l)}, \mathbf{c}_B^{(l)}) \qquad (3)$$

All reported correlations are **pair-level centroid_asymmetry** unless explicitly stated otherwise.

Layer-wise Correlation. We compute Pearson correlation between asymmetry and output preference:

$$r(l) = \text{corr}(\text{asym}^{(l)}, \text{pref}) \qquad (4)$$

3.4 Statistical Analysis

All correlations are computed with bootstrap confidence intervals ($n = 10,000$ resamples, 95% CI). We report significance only when CIs exclude zero.

Methodological Note on Aggregation. Preliminary analysis showed that aggregating metrics at the category level ($n = 6$) masked the phase structure due to intra-category variance, creating a Simpson's Paradox effect. Specifically, category-level correlations showed no significant pattern, while pair-level analysis ($n = 230$) revealed the layer-wise sign change. Our pair-level bootstrapping was necessary to resolve the true signal. We report category-level results in the Appendix for transparency.

3.5 Definition

Definition 1 (Phase-Structured Dynamics). We call the embedding-output relationship *phase-structured* if $r(l)$ shows a **systematic regime change** across depth—for example, a sustained sign change from predominantly positive to predominantly negative, or distinct plateaus with different mean correlations. We do not require both regimes to independently reach significance; rather, we require that the overall pattern shows non-monotonic, depth-dependent structure. Significance is reported per layer separately.

Rationale: This definition accommodates cases where early correlations are weak but positive (Llama) alongside cases with strong positive early correlations (Pythia, Apertus), while excluding models with no discernible regime change.

3.6 Phase Partitions

For comparability across models, we report phase means using a fixed partition:

- Early: Layers 0-8 (or 0-6 for Gemma's 18 layers)
- Mid: Layers 12-20 (or 8-12 for Gemma)
- Late: Layers 24-32 (or 14-18 for Gemma)

Note: Transition points are estimated per-model and may not align with these fixed boundaries. The fixed partition enables cross-model comparison; per-model transitions are reported in Section 4.3.

4 Results

4.1 Main Finding: Phase-Structured Dynamics

Table 2 summarizes phase means across models. Three of four models exhibit significant late-layer inversion. Gemma-2B represents a boundary condition.

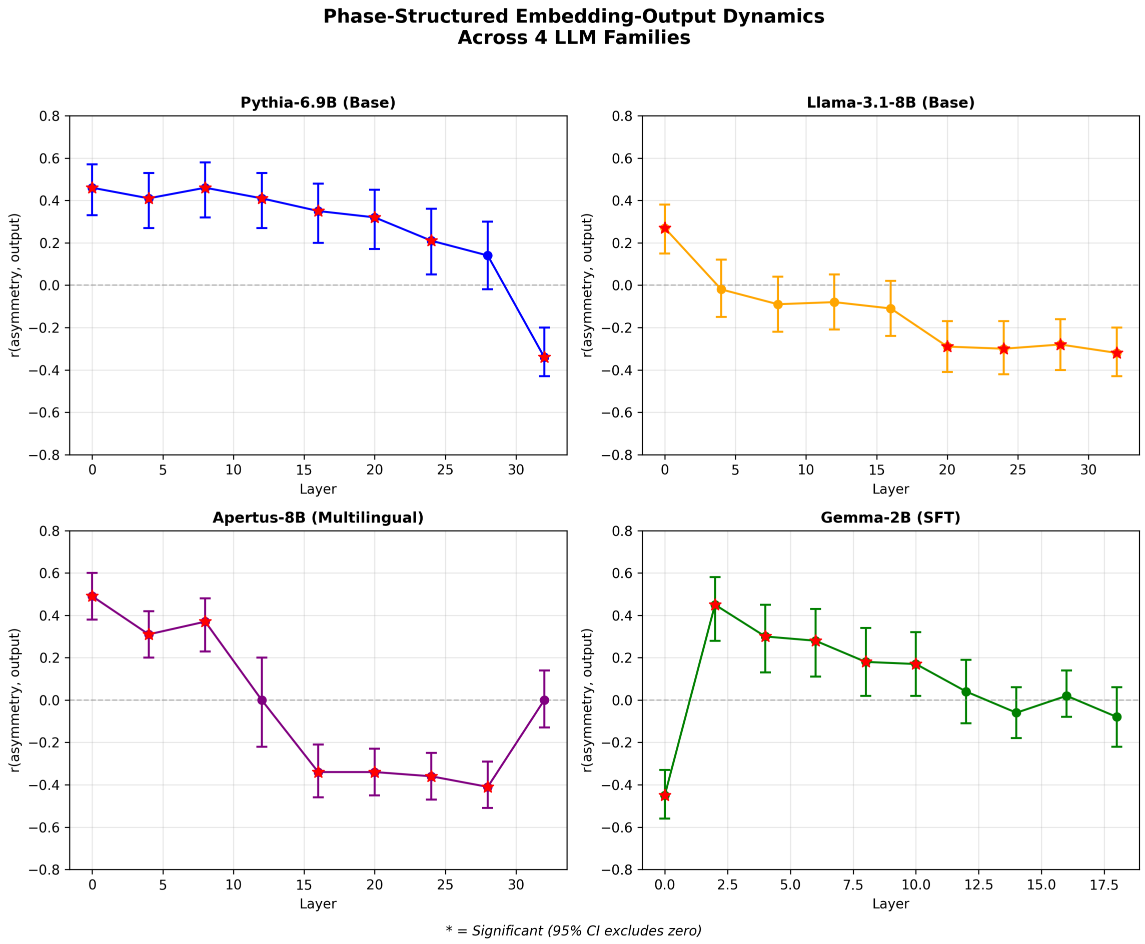

Figure 1: **Phase-Structured Dynamics Across Model Families.** Layer-wise correlation between centroid asymmetry and output preference ($n = 230$ pairs, 10,000 bootstrap resamples). Shaded regions indicate 95% confidence intervals. **Pythia-6.9B** (top-left): Strong positive early correlation with abrupt late inversion. **Llama-3.1-8B** (top-right): Early transition with stable late inversion. **Apertus-8B** (bottom-left): Maximal inversion at Layer 28, not final layer. **Gemma-2B** (bottom-right): Boundary condition with no significant late inversion.

Model	Type	Early	Mid	Late	Phase-Str.?
Pythia-6.9B	Base	+0.44***	+0.36***	**–0.17*****	Yes
Llama-3.1-8B	Base	+0.05	–0.16*	**–0.30*****	Yes
Apertus-8B	Multi.	+0.39***	–0.23***	**–0.25*****	Yes
Gemma-2B	SFT	+0.10	+0.21**	–0.02	No

Table 2: Phase means (Pair-Level Centroid Asymmetry). Early: L0–8, Mid: L12–20, Late: L24–32. Significance: $*p < 0.05$, $**p < 0.01$, $***p < 0.001$. Bold indicates significant late-layer inversion.

4.2 Late-Layer Inversion (Pair-Level Results)

Layer-wise pair-level centroid asymmetry correlations reveal distinct patterns across models (Figure 1). Key observations:

- **Pythia-6.9B:** Strong positive correlation early ($r = +0.46$*** at L0), gradual decline, abrupt inversion at L32 ($r = -0.34$***).
- **Llama-3.1-8B:** Weak positive at L0 ($r = +0.27$***), quick transition to negative, stable late inversion ($r = -0.32$*** at L32).
- **Apertus-8B:** Classic phase structure with maximal inversion at L28 ($r = -0.41$***), not L32.
- **Gemma-2B:** Anomalous L0 ($r = -0.45$***), positive mid-layers, no significant late inversion (L18: $r = -0.08$, CI overlaps zero).

4.3 Transition Point Variation

The layer at which correlation transitions from positive to negative varies substantially:

4.4 Key Observation 1: Apertus Anomaly

In Apertus-8B, maximal embedding-output inversion occurs at Layer 28 ($r = -0.41$, CI: $[-0.51, -0.29]$), while the final Layer 32 returns to near-zero ($r \approx 0.00$, CI: $[-0.13, +0.14]$).

Model	Transition	Characterization
Pythia-6.9B	L28–32	Very late, abrupt
Llama-3.1-8B	L4–8	Early, gradual
Apertus-8B	L12	Mid, sharp
Gemma-2B	—	No clear transition

Table 3: Transition points across models.

One hypothesis is that this pattern reflects a functional decoupling: Layer 28 may serve as the locus of semantic decision commitment, while Layer 32 functions as a projection interface for the unembedding matrix, potentially relaxing geometric constraints to match vocabulary distributions. We emphasize that this interpretation is speculative; causal testing via activation patching remains for future work.

4.5 Key Observation 2: Gemma as Boundary Condition

Gemma-2B (2B parameters, 18 layers, SFT) shows no significant late-layer inversion. All correlations in layers 12-18 have CIs overlapping zero.

Interpretation: This represents a **boundary condition** in our sample where phase structure does not reliably emerge. Possible factors include:

- Scale (2B vs 6.9B-8B)
- Layer count (18 vs 32)
- Training method (SFT vs autoregressive pretraining)

Distinguishing scale from training effects requires further study with matched comparisons (e.g., Gemma-7B if available).

Additionally, Gemma exhibits a unique Layer 0 anomaly: $r = -0.45$*** (significant negative), jumping to $r = +0.45$*** at Layer 2. This indicates fundamentally different input representation dynamics, possibly related to Gemma's distinct embedding initialization or SFT-induced changes.

4.6 Key Observation 3: Template Effects

We tested Llama-3.1-8B-Instruct with and without chat templates:

Condition	Early	Mid	Late
Without template	+0.05	–0.28***	–0.37***
With template	**–0.47*** **	**–0.58*** **	**–0.60*** **

Table 4: Template effect on Llama-3.1-8B-Instruct (Pair-Level). Bold indicates uniformly negative.

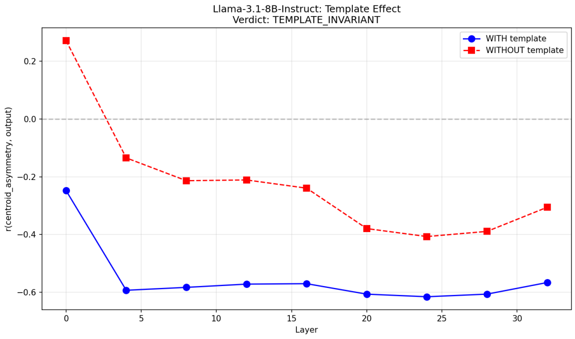

Figure 2: **Template-Induced Processing Regime.** Layer-wise correlations for Llama-3.1-8B-Instruct. **Without chat template** (dashed): Phase-structured dynamics similar to base model. **With chat template** (solid): Uniformly negative correlations across all layers, suggesting a distinct processing regime activated by deployment-format inputs.

Without templates, the instruct model shows phase-structured dynamics similar to the base model. With templates, **all layers show significant negative correlations**.

Interpretation: Instruction tuning **is associated with** a distinct processing regime that appears to be activated by deployment-format inputs (chat templates). Whether this reflects a causal mechanism (templates "switching" modes) or a confound (different input distributions) cannot be determined from correlational analysis alone.

5 Discussion

5.1 What These Results Mean

Our findings support the following interpretation:

> *"Embedding-output relationships in larger base and multilingual LLMs exhibit phase-structured dynamics: early-stage representational geometry often correlates positively with output preference, while decision-relevant inversion emerges at later processing stages, with exact depth being architecture-dependent."*

This is **consistent with** a three-phase processing model:

1. **Early layers (0-8):** Representation building—embeddings cluster by semantic similarity, which in base models correlates positively with output preference.

2. **Mid layers (12-20):** Transition zone—correlation approaches zero or changes sign.

3. **Late layers (24-32):** Decision commitment—compression toward output distribution may invert input-space clustering patterns.

Caveat: We establish correlation, not causation. The mechanistic explanation remains to be tested through causal interventions.

5.2 What These Results Do NOT Mean

We explicitly avoid several interpretations that might be drawn from our findings. First, we do not claim the existence of a discrete "override switch"—late-layer inversion reflects a gradual transition, not a single layer that overwrites earlier computations. Second, our findings concern geometric relationships between embeddings and outputs, not model truthfulness or deception; we make no claims about alignment "killing honesty." Third, phase structure may be functional rather than pathological, potentially reflecting necessary compression for mapping high-dimensional representations to vocabulary distributions. Fourth, transition depth varies substantially by architecture (Layer 4–8 for Llama, Layer 28–32 for Pythia), so there is no universal "tipping point" that applies across models. Finally, while Gemma-2B represents a boundary case where late-layer inversion does not emerge, we cannot determine whether this reflects scale limitations, training methodology, or architectural factors without matched comparisons.

5.3 Connection to Prior Work

This work directly addresses two of the three future research directions outlined in D'Elia (2025): (1) output correlation—we correlate embedding asymmetry with log-probability differences—and (2) layer-wise analysis—we systematically examine how asymmetry evolves across transformer depth. The third direction (downstream task validation) remains for future work.

Our findings converge with Ganesh et al. (2025), who showed through KV-cache manipulation that high-level plans are encoded early while local structure is maintained by final layers. Both studies, using independent methodologies, point toward an early-planning, late-execution processing structure.

The geometric memory hypothesis (Noroozizadeh et al., 2025) explains *why* geometric structure exists (spectral bias); our work shows *what happens* to this geometry across layers during inference.

5.4 Practical Implications

1. **Probing layer selection:** Researchers should be aware that probing early vs. late layers may yield opposite conclusions about model "preferences" or "beliefs."

2. **Interpretability interventions:** If decision commitment depth is model-specific, interventions (activation patching, steering) targeting the "wrong" layer may be ineffective. Our results suggest checking layer-wise correlations before designing interventions.

3. **Template sensitivity:** Safety evaluations using raw text may not reflect deployment behavior for instruction-tuned models. The distinct processing regime under templates should be accounted for.

4. **Safety implications:** If early layers encode representations that are transformed (potentially inverted) in later processing, claims about model "honesty" based on early-layer probing may be misleading.

6 Limitations

Dataset. Our 230 pairs may not capture all semantic domains. The dataset is English-only, uses a specific "uniformity" framing, and the A/B structure is an experimental design choice, not a claim about truth. One might hypothesize that early positive correlations simply reflect representational compression of more uniform statements; however, this cannot explain the observed late-layer *inversion*, which reverses the relationship between geometry and output preference.

Models. We test only four model families in the 2B-8B range. Results may differ for larger models ($>$70B) or different architectures (e.g., mixture-of-experts, SSM).

Metric. Mean-pooling may lose positional information. Centroid-based asymmetry is one of many possible geometric metrics. Alternative metrics (e.g., manifold curvature, cluster separation) might reveal different patterns.

Causation. We establish correlation, not causation. The mechanistic explanation for late-layer inversion—whether it reflects compression, decision commitment, or other processes—remains to be determined through causal interventions (activation patching, ablation).

Confounds. The GT-Numeric category contributes disproportionately to observed effects. We mitigate through pair-level analysis, but category-specific patterns exist. The template comparison may conflate format effects with input distribution differences.

Sequence Length. We did not explicitly control for sequence length differences between statement pairs. Embeddings may correlate with length, though pair-level aggregation mitigates systematic length biases inherent to specific categories.

Generalization. Our boundary case (Gemma) is a single small SFT model. Claims about "scale thresholds" or "SFT effects" require additional models at matched scales and training methods.

7 Conclusion

We present evidence that embedding-output dynamics in large autoregressive transformer models exhibit **phase-structured** patterns: early layers show positive correlation between embedding geometry and output preference, while late layers show inversion. This pattern is confirmed in three of four tested model families (Pythia, Llama, Apertus), with the boundary case (Gemma-2B) suggesting scale or training-method dependencies.

Key findings include:

- Late-layer inversion is architecturally robust in larger base models ($r = -0.17$ to -0.41 in 3/4 models)
- Decision commitment depth is architecture-dependent, and may not localize to the final layer (Apertus: max at L28)
- Small SFT models (Gemma-2B) represent a boundary condition where inversion does not reliably emerge
- Chat templates are associated with a distinct processing regime in instruction-tuned models

Our central conclusion:

> *"Decision commitment in large autoregressive transformers is not uniformly localized at the final layer, but emerges as a late-stage inversion whose depth and sharpness depend on scale, supervision, and architecture. Causal mechanisms remain to be established."*

Open questions for future work:

- Causal testing via activation patching: Does intervening at inversion layers affect output more than other layers?
- Scale curve: At what parameter count does phase structure reliably emerge?
- Training dynamics: When during training does phase structure develop?

Code & Data: `https://github.com/buk81/uniformity-asymmetry`

References

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

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, et al. 2023. Pythia: A suite for analyzing large language models across training and scaling. *arXiv preprint arXiv:2304.01373*.

Davide D'Elia. 2025. Uniformity asymmetry: An exploratory metric for detecting representational preferences in LLM embeddings. *Zenodo preprint doi:10.5281/zenodo.18110161*.

Nelson Elhage, Neel Nanda, Catherine Olsson, et al. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*.

Mukkesh Ganesh, Kaushik Iyer, and Arun Baalaaji Sankar Ananthan. 2025. Whose narrative is it anyway? A KV cache manipulation attack. *arXiv preprint arXiv:2511.12752*.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. 2023. Dissecting recall of factual associations in auto-regressive language models. *arXiv preprint arXiv:2304.14767*.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. 2024. The platonic representation hypothesis. *arXiv preprint arXiv:2405.07987*.

Shahriar Noroozizadeh, Vaishnavh Nagarajan, Elan Rosenfeld, and Sanjiv Kumar. 2025. Deep sequence models tend to memorize geometrically; it is unclear why. *arXiv preprint arXiv:2510.26745*.

Long Ouyang, Jeffrey Wu, Xu Jiang, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Gemma Team, Thomas Mesnard, Cassidy Hardin, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

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.

Hugo Touvron, Louis Martin, Kevin Stone, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Andy Zou, Long Phan, Sarah Chen, et al. 2023. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*.