When Does Geometry Emerge from Memorization in Transformers?

Anonymous authors
Paper under double-blind review

Abstract

Transformer models often show structured internal representations on relational tasks, which are often interpreted as geometric organization. Prior work documents such structure via visualization or performance-based analyses, but does not isolate whether perfect memorization alone yields geometric representations. Here, we conduct a controlled study using synthetic relational worlds defined by canonical graph topologies, explicitly training Transformers to perfectly memorize relational structure without imposing constraints that favor or discourage geometry, and examining whether geometric representations arise as a consequence.

Across chains, cycles, regular graphs, and star graphs, models achieve perfect memorization accuracy while internal embeddings do not systematically preserve either global distances or local neighborhoods, indicating reliance on non-geometric, index-based representations. By probing embeddings against shortest-path distance using rank consistency and neighborhood preservation metrics, we show that memorization alone places no requirement on metric organization. Recoverable geometric structure emerges only when the task objective, together with the relational topology, sufficiently constrains node interchangeability, reducing the space of symmetry-equivalent memorization solutions.

Our results show that perfect memorization does not imply emergent geometric structure, and characterize the conditions under which structure arises in learned embeddings.

1 Introduction

Transformer models are frequently observed to exhibit structured internal representations, such as linear arrangements or clustered layouts, which are often interpreted as evidence that the model has learned a geometric representation of underlying relations (Mikolov et al., 2013b; Hewitt & Manning, 2019). Such observations have motivated the view that geometric organization may emerge naturally during training, and are commonly taken as evidence of relational structure encoded in neural representations (Petroni et al., 2019).

At the same time, achieving perfect accuracy on a relational task does not require preserving distances, neighborhoods, or any notion of metric structure. Prior work has shown that neural models can achieve perfect memorization through arbitrary internal coding schemes, such as index-based or identity-based representations, without analyzing whether such solutions induce any coherent relational or geometric structure (Zhang et al., 2017; Arpit et al., 2017). Transformer architectures in particular admit such associative or lookup-style solutions (Geva et al., 2021), making high task performance alone insufficient to diagnose whether learned representations are genuinely geometric.

This ambiguity complicates the interpretation of probe-based geometric analyses of learned representations: apparent structure may arise even when representations are not uniquely identifiable or causally required for task performance (Hewitt et al., 2021; Elazar et al., 2021). In particular, low-dimensional visualizations from techniques such as UMAP can appear highly structured despite substantial distortions in the underlying metric, and therefore cannot, by themselves, establish meaningful geometric organization (McInnes et al.,

2018). Accordingly, we adopt an operational definition of geometric organization based on consistency between distances in representation space and shortest-path distances in the underlying relational graph.

We study the relationship between memorization, internal geometric organization, and representational stability in a controlled setting. Transformer models are trained from scratch on synthetic relational graph worlds under objectives that permit perfect memorization without requiring generalization, allowing us to isolate when geometric structure is forced by the task rather than incidental. We show that memorization alone does not reliably induce geometric organization, and that stable geometry emerges only when the task objective and relational topology sufficiently constrain node interchangeability, reducing the space of symmetry-equivalent memorization solutions.

2 Methodology

2.1 Objective and Scope

We investigate whether perfect memorization of relational structure necessarily induces a geometric internal representation. To isolate this question, we study Transformer models trained under a pure memorization regime, where the learning objective can be satisfied without compositional reasoning, generalization, or abstraction. Our analysis focuses on the geometry of internal representations rather than task performance alone.

All experiments are conducted in closed synthetic environments, allowing us to disentangle the effects of topology, model capacity, and embedding dimensionality without confounds from linguistic priors or pretraining.

2.2 Why Graphs Rather Than Lexical Data

We analyze synthetic graph-structured data rather than lexical or natural language inputs to achieve precise experimental control. Lexical data introduce confounding factors such as semantics, frequency effects, distributional similarity, and pretraining priors, which obscure the relationship between memorization and internal representation.

Graphs provide an explicit relational structure with a well-defined notion of distance, symmetry, and topology. This allows direct comparison between internal embedding geometry and ground-truth relational distances, enabling us to attribute observed representational effects to structural properties of the task rather than linguistic artifacts.

2.3 Relational Graph Worlds

Each experiment is defined by a graph $G = (V, E)$, where nodes represent discrete entities and edges represent valid successor relations. Given a node $u \in V$, the model is trained to predict one of its successors $v \in \mathcal{N}(u)$, where $\mathcal{N}(u)$ denotes the set of valid successors of u. For chain, cycle, star, and 3-regular graphs, successor relations are directed, ensuring that successor prediction is unambiguous and reducing additional symmetry-based degeneracies; for Watts–Strogatz graphs, edges are undirected and represented using bidirectional successor relations.

Throughout this work, we use *symmetry* to refer specifically to automorphisms of the underlying graph. Graph automorphisms determine when nodes are structurally interchangeable under relabeling. However, node interchangeability under the training objective may also arise from task-induced equivalence, even in graphs with limited formal symmetry. In both cases, large equivalence classes reduce the identifiability of metric structure, permitting multiple functionally correct but geometrically distinct representations. We consider several canonical graph families, each selected to probe a distinct structural property relevant to the emergence of geometric representations.

- **Chain.** A directed path with a unique successor per node. Although the graph has a clear linear ordering, the successor-prediction objective does not require the model to represent this ordering explicitly, allowing multiple functionally equivalent encodings that need not preserve metric structure.
- **Cycle.** A closed loop with uniform degree and rotational symmetry, making all nodes structurally interchangeable and globally ambiguous.
- **Star.** A hub-and-spoke topology with one central node and many structurally identical leaves, yielding shallow and degenerate graph distances.
- **3-Regular Graph.** A graph where each node has degree three. The topology is locally uniform but does not impose a unique global ordering, reducing symmetry without introducing privileged nodes.
- **Watts–Strogatz Small-World Graph.** A graph combining local regularity with randomly rewired long-range connections, introducing multiple length scales and partial symmetry breaking.

Unless otherwise stated, graphs have $|V| = 60$ nodes; we also test $|V| \in 15, 150$ to assess finite-size and scaling effects.

2.4 Model Architecture and Training Regime

We use a decoder-only Transformer trained from scratch, without pretraining or auxiliary objectives. Each node u is mapped to a unique discrete token. Given a single node token the model predicts a distribution over node tokens for valid successors. Although we instantiate a decoder-only Transformer, inputs consist of a single token, so the training objective does not engage sequence-level dependencies or multi-step autoregressive structure. To verify that our findings are not artifacts of the decoder-only architecture, we replicate our core experiments using an encoder-decoder framework, which yields identical conclusions(Appendix C).

Models are trained with cross-entropy (CE) loss for successor prediction, which permits assigning high probability to a single valid successor without enforcing uniformity across neighbors. We also ran a small-scale comparison using a $1/k$ uniform target over neighbors (where $k = |N(u)|$) and observed no qualitative change in the resulting internal geometry. We report CE-based results throughout, as it is the standard and least-constraining objective for this setting.

Training uses a fixed number of epochs sufficient to reach stable convergence and near-perfect set accuracy ($\geq 99.9\%$) across graph families. In deterministic topologies such as chains and cycles, the cross-entropy loss approaches zero. In topologies with multiple valid successors, loss stabilizes at a non-zero value determined by the successor distribution, approaching but not necessarily reaching the uniform-entropy bound, as the training objective does not enforce uniformity across valid successors.

2.5 Embedding Dimension and Graph Size Sweeps

To disentangle geometric effects from representational capacity, we vary both embedding dimensionality and graph size.

Embedding dimension d is varied (e.g., $32, 64, 128, 256$), testing whether geometric organization emerges from increased capacity. Graph size $|V|$ is varied across $15, 60$, and 150 nodes, probing whether larger relational worlds constrain representations more strongly or permit non-geometric memorization strategies to persist.

All other architectural and optimization settings are held fixed across these sweeps.

2.6 Verification of Memorization

To verify that models successfully memorize the relational structure, we compute *set accuracy*. For each node u, a prediction is counted as correct if the model's argmax prediction lies in $\mathcal{N}(u)$. Across all reported settings, models achieve set accuracy above 99.9%, ensuring that differences in internal representations are not attributable to task failure.

2.7 Geometric Probing of Internal Representations

To assess whether memorization corresponds to geometric organization, we analyze frozen hidden representations at every layer. Let $d_G(u, v)$ denote the shortest-path distance between nodes u and v in the graph, and let $d_E(u, v) = \|e_u - e_v\|_2$ denote the Euclidean distance between their embedding vectors. Graph distances are computed using shortest paths; node pairs that are unreachable under the graph are excluded from all distance-based probes.

Spearman Rank Correlation.

$$\rho = \text{Spearman}\big(d_G(u, v), d_E(u, v)\big) \tag{1}$$

This measures whether embedding distances preserve the ordering of graph distances. High values indicate that closer nodes in the graph tend to be closer in the embedding space, even if absolute scaling is distorted.

Coefficient of Determination (R^2).

$$R^2 = 1 - \frac{\sum (d_E - \hat{d}_E)^2}{\sum (d_E - \bar{d}_E)^2}, \quad \hat{d}_E = a\, d_G + b \tag{2}$$

Here, $\hat{d}_E$ denotes the *least-squares linear fit* of embedding distances $d_E(u, v)$ to graph shortest-path distances $d_G(u, v)$ over all node pairs (u, v), and $\bar{d}_E$ denotes the mean embedding distance. This measures how well graph distances linearly predict embedding distances; high values indicate proportional alignment consistent with a global metric embedding.

Metric Distortion.

$$\text{Distortion} = \frac{\max_{u \neq v} \frac{d_E(u,v)}{d_G(u,v)}}{\min_{u \neq v} \frac{d_E(u,v)}{d_G(u,v)}} \tag{3}$$

This quantifies the uniformity of distance scaling in a worst-case sense. Low distortion indicates the absence of extreme pairwise scaling deviations, while high distortion reflects uneven stretching or compression.

Expansion Variance.

$$\text{Var}\left(\frac{d_E(u, v)}{d_G(u, v)}\right) \tag{4}$$

Expansion variance measures how uniformly graph distances are scaled in embedding space. If a coherent global metric is recovered, distances at different graph scales are expanded by similar factors, resulting in relatively low variance. Topologies that exhibit asymmetry, boundary effects, or anisotropy can exhibit non-uniform scaling even when geometry is well formed. In contrast, graphs with degenerate distance structure can produce artificially low variance because many node pairs share identical graph distances.

Local Isometry (1-Hop Stretch).

$$\frac{1}{|V|} \sum_{u \in V} \frac{1}{|\mathcal{N}(u)|} \sum_{v \in \mathcal{N}(u)} d_E(u, v) \tag{5}$$

Local isometry measures how consistently immediate graph neighborhoods are represented in embedding space. When geometry is well formed, neighboring nodes are embedded at comparable distances, indicating stable local structure. Expected values depend on topology, with shallow neighborhoods yielding smaller values and richer local connectivity producing larger values. This metric is therefore interpreted relative to graph structure rather than as an absolute criterion.

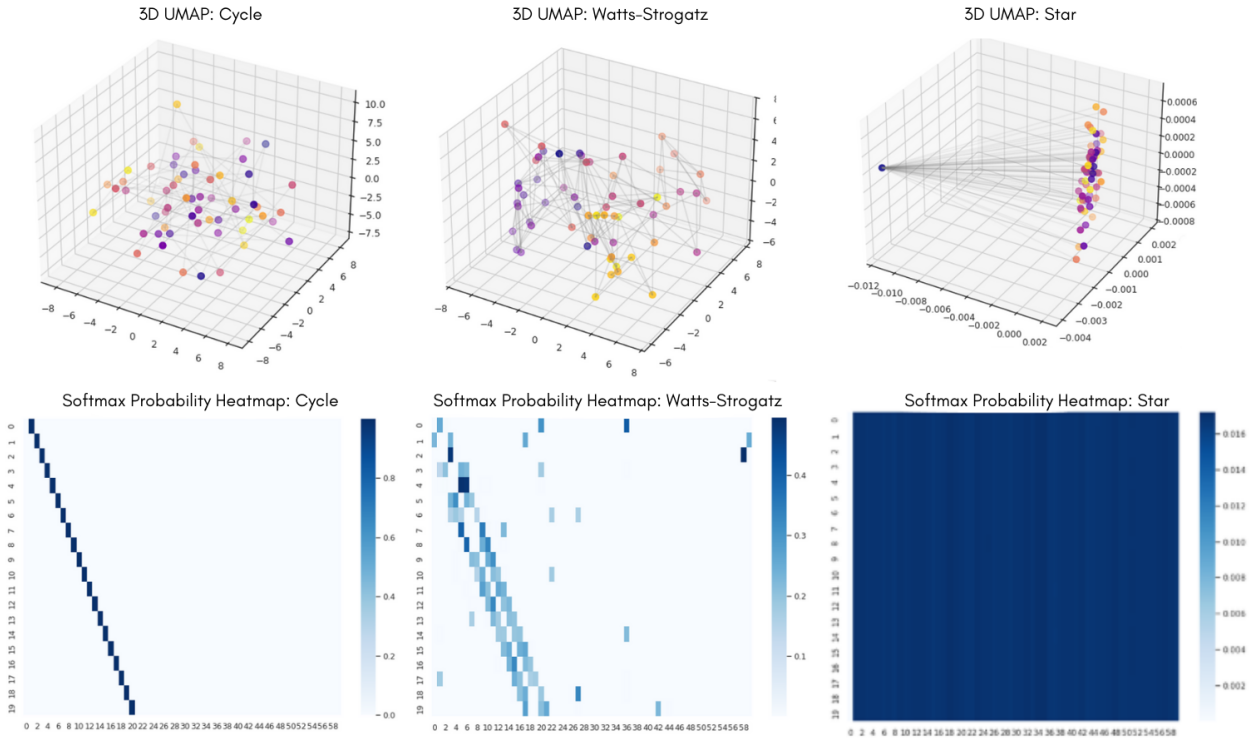

Figure 1: **Geometry vs. memorization across graph topologies.** Top: 3D projections of learned node embeddings. Bottom: successor probability heatmaps. While the model perfectly memorizes local transitions in all cases (bottom),partially recoverable geometric organization emerges only for topologies with sufficient asymmetry (e.g., Watts–Strogatz), and collapses under hub–spoke structure (Star).

2.8 Interpretation of Geometric Outcomes

A model is considered to exhibit *recoverable geometric organization* if embedding distances reflect graph distances under our probes, as indicated by high Spearman (ρ), high R^2, controlled distortion, and *topology-consistent* expansion variance. Local isometry is used as a complementary measure of neighborhood coherence and interpreted comparatively across graph families. Conversely, a model is considered non-geometric if it achieves perfect memorization while failing to preserve globally or locally consistent metric structure.

3 Results

We evaluate whether geometric structure necessarily emerges from memorization by analyzing internal representations of decoder-only Transformers trained on synthetic relational graphs. Results are reported for graphs with $n = 60$ nodes and embedding dimension $d = 128$, a setting that is large enough to admit non-trivial relational structure while remaining small enough to allow detailed geometric analysis; additional experiments across node counts and embedding dimensions are provided in the Appendix.

3.1 Perfect Memorization Does Not Imply Geometry

Despite identical training objectives and perfect memorization accuracy across all graph families, internal representations differ in their geometric properties. Final-layer geometric probe results reveal that memorization alone is insufficient to guarantee the emergence of a coherent metric structure.

Graph	Spearman ρ	R^2	Distort.	Exp. Var	Loc. Iso
Cycle	0.030 ± 0.027	0.001 ± 0.001	70.90 ± 5.23	16.20 ± 1.09	20.35 ± 0.72
Chain	0.028 ± 0.031	0.001 ± 0.002	74.80 ± 7.04	15.71 ± 1.72	19.98 ± 1.15
3-Regular	0.359 ± 0.094	0.103 ± 0.043	5.11 ± 0.65	18.17 ± 1.13	14.23 ± 0.61
Watts–Strogatz	0.673 ± 0.027	0.502 ± 0.030	8.63 ± 0.59	5.37 ± 0.21	12.11 ± 0.18
Star	N/A*	N/A*	6.18 ± 1.57	0.01 ± 0.00	0.18 ± 0.07

Table 1: Aggregate final-layer geometric probe results and functional stability across fifteen random seeds ($n = 60, d = 128$). Results report Mean ± SD. *Spearman ρ and R^2 are mathematically undefined for Star graphs due to the lack of variance in ground-truth shortest-path, rendering correlation and linear regression denominators zero.

3.2 Graphs with High Node Interchangeability Collapse to Non-Geometric Memorization

Quantitative probe results for all graph families are reported in Table 1.

Chain and Cycle. For both the chain and cycle graphs, Spearman rank correlation between embedding distances and graph distances remains near zero, and distortion values are extremely high.Elevated expansion variance in the absence of rank alignment indicates non-uniform scaling unrelated to graph distance, consistent with projection artifacts rather than meaningful geometry.

As shown in Figure 1 (column 1), 3D projections of cycle form scattered point clouds without global ordering. Nevertheless, the corresponding softmax probability distributions are sharply peaked, confirming that successor prediction is memorized precisely. These results indicate that relational worlds with high node interchangeability permit index-based or lookup-style representations that do not require geometry.

3.3 Partial symmetry-breaking: fragile / unstable geometry

3-Regular Graph. The 3-regular graph introduces limited asymmetry while preserving uniform degree. Embeddings show moderate rank correlation, indicating partial alignment between embedding and graph distances. While distortion shows no extreme failures, expansion variance and local isometry are inconsistent with the global structure supported by this topology, reflecting non-uniform scaling and unstable neighborhood structure. As a result, global ordering remains weak. These results indicate that limited asymmetry permits weak but incomplete geometric alignment, insufficient to support a stable global metric embedding.

3.4 Stronger symmetry-breaking: recoverable geometry

Watts–Strogatz Graph. The Watts–Strogatz small-world graph reduces global symmetry while preserving local connectivity. Embeddings achieve higher rank correlation and substantially higher R^2, indicating stronger recovery of global distance ordering, with moderate distortion and relatively low, topology-consistent expansion variance. Embeddings in Figure 1 (column 2) form a coherent low-dimensional structure in which graph distances are partially recoverable, and successor probabilities assign mass to multiple valid successors due to the presence of multiple admissible paths. Together, these results show that stronger symmetry breaking constrains memorization sufficiently for geometry to become recoverable under our probes.

3.5 Star Graph as a Structural Control

The star graph is a control case in which geometric structure is degenerate and imposed by the graph topology. Rank-based metrics are undefined due to the collapse of graph distances, while distortion, local isometry, and expansion variance take trivially non-informative values due to distance degeneracy.

As shown in Figure 1 (column 3), the hub–spoke topology enforces a simple geometric separation between the hub and spokes. This confirms that the probes distinguish geometry imposed by topology from geometric organization that arises under relational constraints.

3.6 Summary

Across all experiments, we observe a consistent pattern: memorization is compatible with multiple representational strategies, only some of which are geometric. Relational worlds with large symmetry-equivalence classes permit non-geometric solutions, partial symmetry-breaking yields fragile geometry, and sufficiently constrained topologies enable stable recoverable geometric organization. Together, these results show that geometric structure is not an automatic consequence of memorization, but instead depends critically on the topological and symmetry constraints of the underlying relational environment.

4 Discussion

4.1 Memorization Permits Multiple Representational Realizations

Our experiments show that perfect memorization does not uniquely determine the geometric organization of internal representations. Across all graph families, models achieve near-perfect set accuracy and approach their task-dependent entropy lower bound, yet exhibit markedly different geometric properties when probed. This reflects that the training objective admits multiple internal realizations that are functionally equivalent for memorization.

Whether geometric structure becomes recoverable depends on how relational constraints are jointly imposed by topology and task. Topology-induced constraints arise from structural asymmetries in the graph itself, while task-induced constraints arise from the successor-prediction objective, which may render distinct nodes functionally interchangeable. Graphs such as Chains and Cycles possess well-defined topological structure, but the task objective does not expose this structure, permitting non-geometric memorization solutions. The 3-regular graph introduces partial asymmetry, weakly restricting interchangeability and inducing fragile, locally coherent geometry that does not stabilize into a global metric. Conversely, the Star graph imposes strong topological asymmetry, but the task remains degenerate due to collapsed distance structure, yielding only trivial, topology-imposed geometry. In contrast, the Watts–Strogatz graph introduces persistent structural heterogeneity that, when coupled with the successor-prediction objective, restricts the space of admissible memorization solutions enough for graph distances to become partially recoverable from embedding distances under our probes.

4.2 Symmetry Suppresses Metric Identifiability

Here we formalize symmetry in terms of graph automorphisms, which determine when nodes are interchangeable under the training objective. Graph families such as cycles admit large automorphism groups, while others such as chains exhibit task-induced interchangeability under the successor objective despite limited formal symmetry. From a representational perspective, this implies that multiple embeddings related by permutation or rotation are equally valid solutions to the training objective. As a consequence, there is no unique or preferred metric ordering that must be recovered in embedding space.

This symmetry manifests empirically as near-zero rank correlation between embedding distances and graph distances, along with persistently high distortion. Because successor-preserving relabelings, whether arising from graph automorphisms or from task-induced equivalence, do not affect prediction accuracy, embeddings need not align with graph-theoretic distances.

4.3 Symmetry-Breaking Constrains Representational Degrees of Freedom

Graph topologies that introduce asymmetry or irregular connectivity reduce the space of functionally equivalent representations. When nodes differ in their relational roles, automorphism-induced equivalence classes collapse, and representational solutions must respect these distinctions to satisfy the training objective.

In such settings, embedding distances become more strongly coupled to graph distances, resulting in improved rank consistency and reduced distortion. However, this coupling is conditional rather than automatic. Increasing embedding dimension expands the space of admissible representations and may suppress geometric alignment unless relational asymmetry sufficiently constrains non-metric solutions. This explains why

geometry does not emerge monotonically with capacity, and why scale alone does not guarantee metric structure.

4.4 Capacity, Scale, and Robustness Across n and d

Varying the number of nodes and embedding dimension confirms that the observed behaviors are robust to scale. In symmetric graph families, increasing n or d preserves near-perfect memorization while leaving geometric probes unchanged, indicating that scale does not resolve metric non-identifiability. In contrast, for asymmetric graph families, increased capacity stabilizes geometric organization when topological constraints limit alternative realizations.

These findings indicate that geometric structure arises from the interaction between model capacity and relational constraints, rather than from scale or optimization dynamics alone.

4.5 Boundary Cases and Implications

The star graph provides a boundary case in which geometric structure is imposed directly by topology rather than resolved by the task. Owing to its hub–spoke organization, geometry is trivial and unavoidable and rank-based probes are degenerate by construction. This illustrates that geometric organization may be imposed, recoverable, or absent depending on the relational structure.

More broadly, these findings indicate that geometric analyses of learned representations should be interpreted relative to task identifiability, rather than treated as intrinsic properties of the architecture. When the training objective admits multiple functionally equivalent solutions, the presence or absence of geometric structure alone is insufficient to diagnose how relational information is encoded.

5 Related Work

5.1 Geometric Structure in Transformer Representations

A growing literature reports that Transformer models often exhibit structured or geometric organization in their hidden representations. Early evidence of linear and hierarchical structure in neural embeddings emerged from studies of word representations (Mikolov et al., 2013a; Pennington et al., 2014), followed by analyses showing that contextual Transformers encode syntactic and semantic relations in linearly recoverable form (Hewitt & Manning, 2019; Jawahar et al., 2019). More recent work has studied the geometric structure of transformer representations across layers and tasks (Valeriani et al., 2023).

5.2 Associative Memory, Lookup Tables, and Memorization

A long-standing view models neural memorization as associative or lookup-based storage of co-occurrences, where embeddings act as arbitrary identifiers and relations are stored in weight matrices (Hopfield, 1982). This abstraction has been widely adopted in analyses of Transformer memory (Geva et al., 2021), and is often sufficient to explain behavior on disjoint or atomic facts.

Several works argue that, in the absence of structural constraints, Transformers can implement near-perfect lookup-table solutions (Arpit et al., 2017). Recent empirical studies suggest that increased model capacity primarily improves memorization fidelity rather than inducing structured representations. Our findings align with this view in symmetric relational worlds, while clarifying the conditions under which associative solutions coexist with or give way to geometric ones.

5.3 Probing, Identifiability, and Representation Analysis

Probing methods are commonly used to assess what information is encoded in neural representations (Adi et al., 2017; Conneau et al., 2018). However, multiple works caution that probe success does not imply that a representation intrinsically encodes a property, as probes may exploit incidental correlations or task structure (Hewitt et al., 2021; Ravichander et al., 2021). These observations motivate metric, rank, and local-structure

probes beyond linear separability (Ethayarajh, 2019), alongside information-theoretic alternatives (Pimentel et al., 2020). Our results highlight a limitation of geometric probing analyses commonly used in natural language processing. When the training objective admits multiple functionally equivalent representations, observed geometric structure may be incidental rather than identifiable. Probing results should therefore be interpreted relative to the constraints imposed by the task, rather than assumed to reflect learned relational understanding.

Our approach builds on this perspective by evaluating geometric organization using complementary distance-based probes and by analyzing robustness across scales. We connect geometric probing to task identifiability: when multiple representational realizations are functionally equivalent, probes may reveal the absence of a uniquely recoverable geometry.

5.4 Relational and Graph-Based Learning

Relational reasoning has traditionally been studied using graph neural networks, which impose explicit inductive biases for locality and message passing (Kipf & Welling, 2017; Hamilton et al., 2017). Transformers, despite lacking such biases, have shown strong empirical performance on graph and relational tasks (Dwivedi et al., 2023). Several works explore hybrid or geometry-aware architectures that explicitly encourage metric structure (Chen et al., 2022).

Our work differs from prior work by holding architecture fixed and varying relational topology. This allows us to isolate how symmetry and asymmetry in the task constrain representational geometry, independent of architectural enforcement.

6 Conclusion

We investigated whether geometric organization is a necessary consequence of memorization in Transformer models by analyzing internal representations learned on synthetic relational graph worlds. Across a range of graph topologies, node counts, and embedding dimensions, we show that perfect memorization is compatible with multiple representational realizations, only some of which exhibit coherent geometric structure. Our results identify topology induced symmetry breaking in the task as a central factor governing when geometry is recoverable from the training objective. Relational worlds with large equivalence classes arising from graph symmetry or from task-induced interchangeability admit non-geometric memorization strategies even at scale. Importantly, increasing model capacity alone does not guarantee geometric emergence and may instead expand the space of functionally equivalent, non-metric solutions. These findings clarify the distinction between memorization and representation, and caution against interpreting geometric structure as an intrinsic or automatic property of Transformer architectures. More broadly, our work highlights the role of task identifiability in representation analysis and provides a principled framework for evaluating when geometric organization reflects genuine relational structure rather than functionally equivalent alternatives.

7 Limitations

Our study focuses on controlled synthetic relational environments designed to isolate the representational consequences of memorization. While this setting enables precise analysis of identifiability and geometric structure, it does not capture the full complexity of natural language or multimodal data, where semantic grounding, redundancy, and compositional generalization may introduce additional constraints on representations.

We analyze a fixed Transformer architecture trained under a successor-prediction objective. Different architectural biases, training objectives, or regularization schemes may alter the space of admissible representational solutions. In addition, although we employ multiple complementary geometric probes, all probing analyses depend on the choice of metric and evaluation procedure. Our conclusions therefore concern the existence and stability of *recoverable* geometric structure under these probes, rather than asserting the absence of any latent organization.

References

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. In *International Conference on Learning Representations*, 2017.

Devansh Arpit, Stanislaw Jastrzebski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S. Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, and Simon Lacoste-Julien. A closer look at memorization in deep networks. In *International Conference on Machine Learning*, 2017.

Dexiong Chen, Leslie O'Bray, and Karsten Borgwardt. Structure-aware transformers for graph representation learning. In *International Conference on Machine Learning*, 2022.

Alexis Conneau, Germán Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. What you can cram into a single vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2018.

Vijay Prakash Dwivedi, Chaitanya K. Joshi, Anh Tuan Luu, Thomas Laurent, Yoshua Bengio, and Xavier Bresson. Benchmarking graph neural networks. *Journal of Machine Learning Research*, 2023.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 2021.

Kawin Ethayarajh. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings. In *Proceedings of EMNLP-IJCNLP*, 2019.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In *Empirical Methods in Natural Language Processing*, 2021.

William Hamilton, Rex Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems*, 2017.

John Hewitt and Christopher Manning. A structural probe for finding syntax in word representations. In *Proceedings of NAACL*, 2019.

John Hewitt, Kawin Ethayarajh, Percy Liang, and Christopher D. Manning. Conditional probing: Measuring usable information beyond a single probe. In *Proceedings of NAACL*, 2021.

John Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, 1982.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. What does bert learn about the structure of language? In *Proceedings of ACL*, 2019.

Thomas Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017.

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, 2013a.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of NAACL*, 2013b.

Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing*, 2014.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. Language models as knowledge bases? In *Empirical Methods in Natural Language Processing*, 2019.

Tiago Pimentel, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell. Information-theoretic probing for linguistic structure. In *Proceedings of ACL*, 2020.

Abhilasha Ravichander, Yonatan Belinkov, and Eduard Hovy. Probing the probing paradigm: Does probing accuracy entail task relevance? In *Proceedings of EACL*, 2021.

Lucrezia Valeriani, Diego Doimo, Francesca Cuturello, Alessandro Laio, Alessio Ansuini, and Alberto Cazzaniga. The geometry of hidden representations of large transformer models. *arXiv preprint arXiv:2302.00294*, 2023. Presented as a poster at NeurIPS 2023.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *International Conference on Learning Representations*, 2017.

8 Appendix

A Experimental Setup, Resources, and Reproducibility

This section documents the experimental setup, hyperparameters, computational resources, and implementation details required to reproduce the results in the paper. The goal is to ensure transparency and reproducibility rather than to optimize for computational efficiency.

Model architecture and size. All experiments use a decoder-only Transformer trained from scratch. The architecture follows a GPT-2–style decoder with 8 Transformer layers, each employing multi-head self-attention with 4 attention heads over a d-dimensional embedding space. Embedding dimensionality is varied across experiments ($d \in \{32, 64, 128, 256\}$), and graph size is varied over $|V| \in \{15, 60, 150\}$. Each node in the graph is mapped to a unique discrete token, and the model predicts successor tokens given a single-node input. As a reference configuration, for $d = 128$ and $|V| = 60$ (the primary setting reported in the main paper), the model contains approximately 1.5–1.7 million parameters, depending on bias terms and embedding size.

Training procedure and hyperparameters. Models are trained using the AdamW optimizer with learning rate 1×10^{-3}, weight decay 0.1, and batch size 32. The training objective is cross-entropy loss over successor prediction. All runs are trained for a fixed 10,000 epochs, which is sufficient to reach stable convergence and near-perfect set accuracy ($\geq 99.9\%$) across all graph families. We do not employ early stopping, validation-based model selection, or automated hyperparameter search. Hyperparameter values were chosen based on standard practice and empirically verified to yield stable convergence across graph families and model sizes. Hyperparameters were not tuned on held-out evaluation data.

Graph construction and data generation. Experiments are conducted on synthetic relational graph worlds constructed using standard graph generators. For chain, cycle, star, and 3-regular graphs, successor relations are directed. For Watts–Strogatz graphs, edges are undirected and represented using bidirectional successor relations. Watts–Strogatz graphs are generated with nearest neighbors $k = 4$ and rewiring probability $p = 0.2$. Unless otherwise stated, graphs contain $|V| = 60$ nodes. Shortest-path distances used in geometric probes are computed with respect to the graph definition for each topology, and node pairs that are unreachable under the corresponding graph are excluded from distance-based analyses.

Random seeds and reporting. For the primary configuration reported in the main paper ($d = 128$, $|V| = 60$), all quantitative results are aggregated over 15 independent random seeds and reported as mean $\pm$ standard deviation. For all other configurations (embedding-dimension sweeps, graph-size sweeps, and ablations reported in the Appendix), results are reported from a single representative run. Unless otherwise stated, all reported metrics correspond to the final Transformer layer after convergence.

Computational resources. All experiments were run on Google Colab CPU instances. Training time varied substantially depending on graph topology, graph size, embedding dimension, and random seed. Even the smallest configurations required on the order of an hour per training run, while larger or higher-dimensional settings required multiple hours. The full 15-seed sweep for the $d = 128$, $|V| = 60$ configuration required approximately two days of computation. Across all experiments, including embedding-dimension sweeps, graph-size sweeps, loss-function ablations, and multiple random seeds, the total computational budget is estimated at several days of CPU computation.

Software and implementation. Models are implemented in PyTorch using standard Transformer components. Graph construction and shortest-path computations use NetworkX. UMAP is used exclusively for qualitative visualization and is not used for quantitative evaluation or for any claims in the paper. All experiments rely on publicly available libraries, and no external codebases were modified. Code to reproduce all experiments will be made available upon acceptance.

B Additional Experimental Results

This appendix provides supplementary experimental results supporting the claims in the main paper. Unless otherwise stated, all values are reported from the *final layer* of the model (Layer 8), reflecting representations at convergence.

All tables use the same geometric probes as in the main paper. Higher Spearman correlation and R^2 indicate stronger geometric alignment with graph shortest-path distance, while lower distortion indicates greater metric uniformity. Expansion variance and local isometry provide complementary diagnostics of low-dimensional structure and neighborhood preservation.

B.1 Embedding Dimension Sweeps (Cross-Entropy Loss)

We examine how embedding dimensionality affects geometric organization when memorization accuracy is held constant. Graph size is fixed at $n = 60$, and all models are trained using cross-entropy loss. Across all settings, models achieve perfect memorization, allowing us to isolate representational geometry from task performance.

Low-dimensional regime ($d = 32$). Results are reported in Table 2. At low dimensionality, representational capacity is limited, restricting the set of embedding configurations that can achieve perfect memorization. In Chain and Cycle graphs, this constraint does not induce geometry the successor-prediction task exposes no global ordering, and nodes remain largely interchangeable. As a result, embeddings fail to reflect graph distances despite the apparent simplicity of these topologies.

In contrast, the 3-regular graph already exhibits measurable geometric alignment at $d = 32$. Although all nodes have identical degree, higher-order connectivity patterns differ, introducing weak but genuine asymmetry. Under limited capacity, these differences partially constrain representations, yielding non-zero rank correlation. However, scaling remains inconsistent, indicating that this alignment is local rather than globally metric.

Watts–Strogatz graphs show clearer geometric structure even at low dimensionality. Long-range shortcuts introduce global asymmetry by differentiating nodes according to path centrality and multi-hop reachability, allowing embeddings to reflect graph distances even under tight capacity constraints.

Intermediate dimensionality ($d = 64$). Results are reported in Table 3. As dimensionality increases, representational capacity expands. For Chain and Cycle graphs, this does not improve geometry: the task still permits index-based solutions, and additional dimensions only increase the space of non-metric representations.

For the 3-regular graph, increased capacity weakens geometric alignment relative to $d = 32$. Though local neighborhoods remain structured, the graph's near-regularity permits many nodes to be represented by

shared functional roles rather than arranged according to global graph distance. The model can satisfy the task using these alternative encodings, leading to reduced rank consistency.

Watts–Strogatz graphs, by contrast, benefit from increased dimensionality at this stage. The presence of shortcuts continues to break symmetry at multiple scales, and additional capacity supports a more stable embedding that better reflects global graph distances.

High-dimensional regime ($d = 256$). Results are reported in Table 4. At high dimensionality, these trends become more pronounced. Chain and Cycle graphs remain non-geometric, as additional capacity further decouples representations from any implicit ordering.

In the 3-regular graph, geometric alignment deteriorates further. Without strong global anchors, the model increasingly exploits high-dimensional freedom to realize symmetry-equivalent memorization strategies. This results in unstable geometry characterized by reduced rank alignment, increased distortion, and noticeable variability across random seeds.

Watts–Strogatz graphs, however, maintain stable geometric organization at higher dimensions. Once shortcut-induced asymmetry establishes a meaningful global structure, additional dimensions neither disrupt nor substantially enhance geometry. Instead, embeddings saturate into a consistent configuration that preserves graph distances under our probes.

Summary. Embedding dimensionality interacts with graph topology in a graph-specific manner. Highly interchangeable structures such as Chains and Cycles remain non-geometric across all dimensions. Graphs with weak but higher-order structural asymmetry, such as 3-regular graphs, exhibit dimension-sensitive and seed-unstable geometry. In contrast, graphs with persistent multi-scale asymmetry, such as Watts–Strogatz graphs, support stable geometric organization across a wide range of embedding dimensions. The star graph forms a distinct boundary case its degenerate hub–spoke distance structure imposes a trivial geometric separation that remains unchanged across embedding dimensions. These results highlight that changes in embedding dimension affect geometry through their interaction with specific structural features, rather than through capacity alone.

B.2 Loss Function Ablation

We next test whether representational collapse is specific to the cross-entropy objective. We compare against a uniform $1/K$ loss that assigns equal probability mass to all valid successors, holding graph size ($n = 60$) and embedding dimension ($d = 128$) fixed. All comparisons are reported for the *final layer* only, after models achieve stable convergence and near-perfect set accuracy. Aggregate geometric probe results are reported in Table 5.

Across all graph families, the choice of loss function does not qualitatively alter whether geometric structure emerges at convergence. Graph families with high node interchangeability (Chain and Cycle) remain non-geometric under both objectives, exhibiting near-zero rank correlation and extremely high distortion despite perfect memorization. This indicates that representational collapse in these settings is not an artifact of the cross-entropy loss, but rather a consequence of the task admitting multiple functionally equivalent, non-metric solutions.

For the 3-regular graph, which introduces partial asymmetry while maintaining uniform degree, the uniform $1/K$ objective does not reliably induce recoverable geometry. In the reported run, rank correlation remains modest and global scaling inconsistency persists, with moderate distortion and high expansion variance. Given the weak and unstable geometry already observed under cross-entropy, these results indicate that partial symmetry alone is insufficient to support a coherent global metric embedding across objectives, and that apparent differences may reflect seed-level variability rather than a systematic effect of the loss function.

In the Watts–Strogatz graph, which strongly breaks global symmetry while preserving structured local connectivity, robust geometric organization emerges under both loss functions. Final-layer rank correlation and R^2 remain high, and distortion remains controlled, demonstrating that sufficiently strong relational

constraints induce geometric structure regardless of the training objective. The $1/K$ loss primarily affects the smoothness of output distributions rather than the presence of geometry itself.

Finally, the star graph exhibits degenerate geometric behavior under both objectives. Rank-based metrics are undefined due to the lack of variance in graph distances, while distortion and expansion variance reflect topology-imposed constraints rather than learned metric organization. Under the uniform $1/K$ objective, local isometry is higher, reflecting the trivial and identical 1-hop neighborhoods imposed by the hub–spoke topology. This does not indicate meaningful geometric structure, but rather highlights that purely local probes become ill-conditioned in settings with extreme distance degeneracy.

Summary. Altering the training objective may modulate the stability of weak geometric effects, but does not determine whether geometric structure emerges. Recoverable geometry at convergence is governed primarily by the topological and symmetry constraints of the underlying graph, rather than by the choice of loss function.

B.3 Effect of Graph Size

Finally, we evaluate whether representational collapse is a finite-size artifact by varying graph size under cross-entropy training.

Small graphs. Results for smaller graphs are reported in Table 6. At small graph sizes ($n = 15$), geometric alignment emerges selectively depending on the degree of relational identifiability imposed by graph topology. Graph families with limited node interchangeability exhibit measurable geometric alignment, while graph families with high interchangeability continue to admit non-geometric memorization solutions (Table 6). In this small-n regime, each relational constraint influences a large fraction of node pairs, making global geometric probes particularly sensitive to weak symmetry-breaking effects.

At small n, partial asymmetries introduced by graph topology constrain the space of admissible representations more strongly, as fewer functionally equivalent embeddings satisfy the training objective. This finite identifiability regime amplifies weak topological constraints that would otherwise be diluted at larger scales. As a result, even incomplete relational structure can induce measurable rank alignment between embedding distances and graph distances.

This effect is most evident in the 3-regular graph. Although all nodes share identical degree, higher-order connectivity patterns differ substantially at $n = 15$, including variations in cycle participation, neighborhood overlap, and multi-hop reachability. These differences reduce node interchangeability and limit the viability of purely index-based memorization strategies. Consequently, the model exhibits relatively high Spearman rank correlation, indicating partial geometric alignment. However this geometry remains fragile, distortion and expansion variance remain elevated and not topology-consistent , reflecting inconsistent global scaling and the absence of a coherent metric embedding despite non-zero rank alignment.

In contrast, the Watts–Strogatz graph does not exhibit strong geometric alignment at $n = 15$. While the topology introduces long-range shortcuts that break global symmetry, at small n these shortcuts are too sparse to establish a stable global ordering of nodes. Local ring structure dominates, and many nodes remain functionally similar under the successor prediction objective. As a result, the model can satisfy the task using localized memorization strategies that do not require embedding distances to reflect global graph distances. This explains the low rank correlation observed despite relatively bounded distortion, indicating that low distortion alone is insufficient to guarantee meaningful geometric organization.

Highly symmetric graph families such as chains and cycles continue to admit non-geometric memorization strategies even at small scale. In these settings, nodes remain functionally interchangeable under the training objective, allowing the model to achieve perfect memorization without encoding relational distances. Accordingly, rank correlation remains near zero and distortion remains high, indicating representational collapse rather than metric structure.

The star graph represents a distinct boundary case. At $n = 15$, graph-theoretic distances are degenerate, consisting primarily of hub–spoke relations with no meaningful ordering among spoke nodes. Ratio-based

geometric probes such as distortion are therefore ill-conditioned in this regime individual spoke placements exert disproportionate influence on extreme distance ratios, leading to elevated distortion values. Importantly, this does not reflect a failure to learn task-relevant structure. The model learns a task-aligned geometric separation between the hub and the set of spokes that is sufficient for perfect successor prediction, while distances among spokes remain unconstrained and unidentifiable.

Together, these observations demonstrate that geometric structure does not arise from small scale alone. Rather, small-n regimes amplify the effect of topological constraints, revealing partial geometry when relational asymmetries restrict the solution space, and exposing metric degeneracies when relational structure remains underdetermined.

Large graphs. Results for larger graphs are reported in Table 7. At larger graph sizes ($n = 150$), the behavior of geometric probes diverges sharply across graph families, reflecting how each topology interacts with scale under the memorization objective.

For the 3-regular graph, the partial geometric alignment observed at smaller n disappears entirely. Although node degree remains fixed, increasing graph size renders the topology progressively more homogeneous most nodes occupy statistically similar positions with respect to local neighborhoods and multi-hop connectivity. As n grows, these similarities dominate the task signal, making many nodes functionally indistinguishable from the model's perspective. The successor prediction objective can therefore be satisfied by grouping nodes into a small number of role-based equivalence classes rather than organizing them in a globally consistent metric space. This leads to representational collapse, characterized by near-zero rank correlation and elevated distortion despite perfect memorization.

Chain and cycle graphs exhibit a different failure mode. While these topologies possess a well-defined global metric structure, the successor prediction task does not require the model to preserve long-range distances as graph size increases. As n grows, the majority of training signal is dominated by local transitions, and errors in global ordering become increasingly irrelevant to the loss. Consequently, embeddings need only preserve local adjacency rather than coherent global geometry, resulting in vanishing rank correlation and severe distortion at scale.

In contrast, Watts–Strogatz graphs maintain stable geometric alignment even at large n. Although locally near-regular, this topology introduces persistent structural heterogeneity through long-range shortcuts. As the graph grows, differences in shortcut participation, features prevent large equivalence classes of interchangeable nodes from forming, forcing the model to preserve global distance relationships in order to satisfy the training objective. As a result, geometric alignment remains stable, with moderate rank correlation and controlled distortion.

Finally, star graphs represent a degenerate boundary case. The presence of a single hub imposes an explicit and unavoidable geometric separation between the center and the leaves. This structure is trivially imposed by the topology and does not depend on learning or scale. After excluding unreachable node pairs, star graphs admit only a single non-trivial distance class, rendering rank-based geometric probes degenerate by construction. Accordingly, star graphs are excluded from correlation-based analysis.

Together, these results show that increasing graph size does not inherently promote geometric organization. Instead, scale interacts with topology by either diluting or preserving the structural signals required for geometric identifiability. Topologies whose relational constraints remain informative as n grows continue to support recoverable geometry, while those that become functionally homogeneous or locally sufficient under the task objective admit non-geometric memorization solutions.

Summary. Across graph families and scales, increasing graph size alone does not induce geometric organization. When topological constraints are weak or homogeneous, perfect memorization is possible without distance-aware representations, leading to probe collapse. In contrast, topologies with persistent structural heterogeneity support recoverable geometry. These results show that geometric structure is governed by topological identifiability rather than graph size or model capacity.

Graph	Spearman	R^2	Distortion	Exp. Var.	Local Iso.
CHAIN	0.0657	0.0017	58.67	8.2647	14.3473
CYCLE	0.1342	0.0164	65.02	7.0019	13.2222
3-REG	0.5257	0.1569	7.20	16.1035	12.8129
WS	0.5685	0.3748	7.13	4.5539	11.2405
STAR	–	–	1.13	0.0000	0.0985

Table 2: Geometry diagnostics for $n = 60$, $d = 32$ using the final layer (Layer 8).

B.4 Qualitative Visualization of Learned Representations

To complement the quantitative geometric probes reported above, we provide qualitative visualizations of learned representations for each graph family. These visualizations are intended as illustrative diagnostics rather than primary evidence, as low-dimensional projections may distort global geometry. Accordingly, all claims in the paper are grounded in the distance-based probes reported in the main text and appendix tables.

Each dashboard visualizes (i) a 2D projection of node embeddings, (ii) a 3D embedding visualization, (iii) the underlying relational graph for a representative node, and (iv) the corresponding softmax successor probability distribution.All visualizations are generated from graphs with 60 nodes, using embeddings of dimension 128.

C Architecture Independence: Encoder-Decoder Transformers

In the main text, our analyses focus on decoder-only Transformers to align with common architectures used in autoregressive language modeling. To verify that the observed relationship between memorization, topological symmetry, and geometric emergence is a fundamental property of the learning objective rather than an architectural artifact, we replicate our core experiments using an encoder-decoder framework. We adopt a standard encoder-decoder Transformer trained from scratch under the identical successor-prediction objective. To maintain capacity comparability with the decoder-only experiments—which utilized 8 layers—we configure the encoder-decoder model to have 4 encoder layers and 4 decoder layers, yielding an equivalent total depth. The embedding dimension is fixed to $d = 128$ with 4 attention heads, and the graph size to $n = 60$. As with the primarily analyzed models, the encoder-decoder configurations achieve near-perfect memorization (set accuracy $\geq 99.9\%$) across all tested graph families.

We probe the internal representations to determine whether introducing an encoder-decoder architecture alters geometric recoverability. Our findings demonstrate that the representations exhibit the precise pattern of geometry emergence observed in decoder-only models (Table 8). In highly symmetric topologies, such as chains and cycles, the model does not develop a coherent global metric. The learned embeddings exhibit near-zero rank correlation with graph shortest-path distances and high distortion, confirming that high node interchangeability permits non-geometric memorization regardless of the encoder-decoder architecture. Under partial symmetry-breaking, such as in 3-regular graphs, the representation space shows only weak and fragile geometric alignment, characterized by inconsistent global scaling.

Conversely, under the strong symmetry-breaking of Watts–Strogatz small-world graphs, coherent and stable geometric structure robustly emerges within the model's internal representations. In the boundary case of star graphs, the geometry degenerates into the trivial, topology-dictated hub-and-spoke separation, confirming it as an imposed limit rather than a learned spatial organization. These results establish that the selective emergence of geometric structure is architecture-agnostic. The non-identifiability of metric structure in symmetric graphs is a direct mathematical consequence of how the task objective interacts with relational topology, rather than an artifact of causal masking or decoder-only processing. Consequently, the conclusion that geometric structure depends primarily on symmetry constraints applies broadly across Transformer architectures.

Graph	Spearman	R^2	Distortion	Exp. Var.	Local Iso.
CHAIN	0.0105	0.0000	88.39	10.4470	16.2047
CYCLE	0.0413	0.0001	60.95	11.2844	16.9093
3-REG	0.3585	0.0847	7.69	21.6430	14.1518
WS	0.6077	0.4043	8.51	5.9267	12.8957
STAR	–	–	1.14	0.0000	0.0197

Table 3: Geometry diagnostics for $n = 60$, $d = 64$ using the final layer (Layer 8).

Graph	Spearman	R^2	Distortion	Exp. Var.	Local Iso.
CHAIN	0.0324	0.0006	170.98	22.3411	23.7879
CYCLE	0.0339	0.0012	94.06	26.9706	26.3165
3-REG	0.2052	0.0099	16.93	22.2633	15.0140
WS	0.6682	0.4634	8.17	5.4428	12.3224
STAR	–	–	2.65	0.0000	0.0192

Table 4: Geometry diagnostics for $n = 60$, $d = 256$ using the final layer (Layer 8).

Graph	Spearman	R^2	Distortion	Exp. Var.	Local Iso.
CHAIN	-0.0361	0.0017	83.13	15.6146	20.0585
CYCLE	0.0596	0.0032	66.41	14.7997	19.6616
3-REG	0.2740	0.0378	15.05	24.2272	15.8122
WS	0.4994	0.3204	9.01	8.7537	15.4345
STAR	–	–	1.05	0.000	0.9402

Table 5: Geometry diagnostics using uniform $1/K$ loss ($n = 60$, $d = 128$), final layer (Layer 8).

Graph	Spearman	R^2	Distortion	Exp. Var.	Local Iso.
CHAIN	-0.2126	0.0042	16.70	12.7608	12.0401
CYCLE	0.1841	0.0477	14.96	10.3763	11.0858
3-REG	0.5590	0.1436	9.42	10.4900	8.4305
WS	0.0563	0.0032	4.96	9.9701	10.6380
STAR	–	–	16.01	1.1222	0.6594

Table 6: Geometry diagnostics for small graphs ($n = 15$, $d = 32$), final layer (Layer 8).

Graph	Spearman	R^2	Distortion	Exp. Var.	Local Iso.
CHAIN	0.0053	0.0000	201.75	5.6996	17.2706
CYCLE	0.0384	0.0014	201.44	6.2304	18.1359
3-REG	-0.0265	0.0005	28.03	23.8432	15.1244
WS	0.5383	0.3879	10.51	5.3564	16.1860
STAR	–	–	1.25	0.0000	0.0046

Table 7: Geometry diagnostics for large graphs ($n = 150$, $d = 384$), final layer (Layer 8).

Graph	Spearman	R^2	Distortion	Exp. Var.	Local Iso.
CHAIN	0.0184	0.0001	80.68	74.6450	44.8407
CYCLE	0.0066	0.0000	81.20	75.3669	45.2177
3-REG	0.2006	0.0075	6.67	142.4757	40.6776
WS	0.5950	0.4106	8.97	64.9011	41.4645
STAR	–	–	320.75	8.2636	6.0476

Table 8: Geometry diagnostics for encoder-decoder models ($n = 60, d = 128$), final layer (Decoder Layer 4).

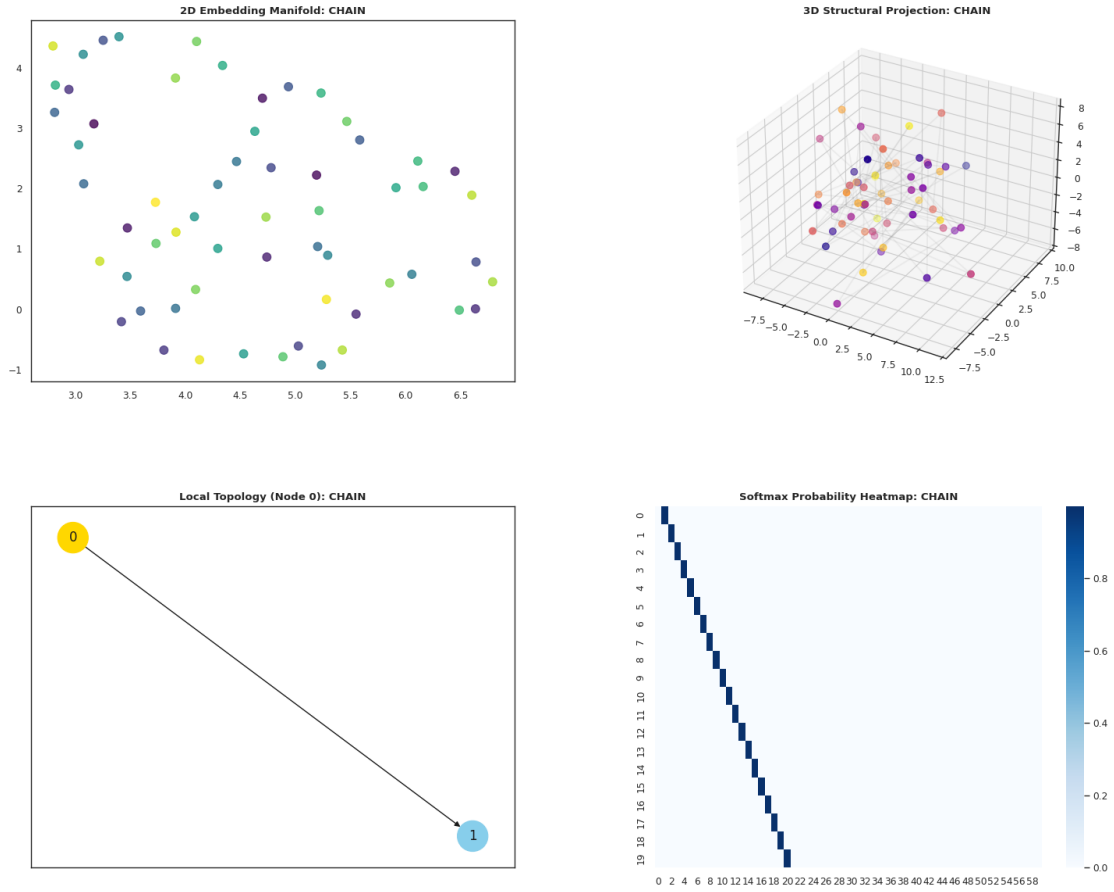

Figure 2: **Chain graph: memorization without geometric structure.** Embeddings do not exhibit a coherent global metric despite perfect task performance.

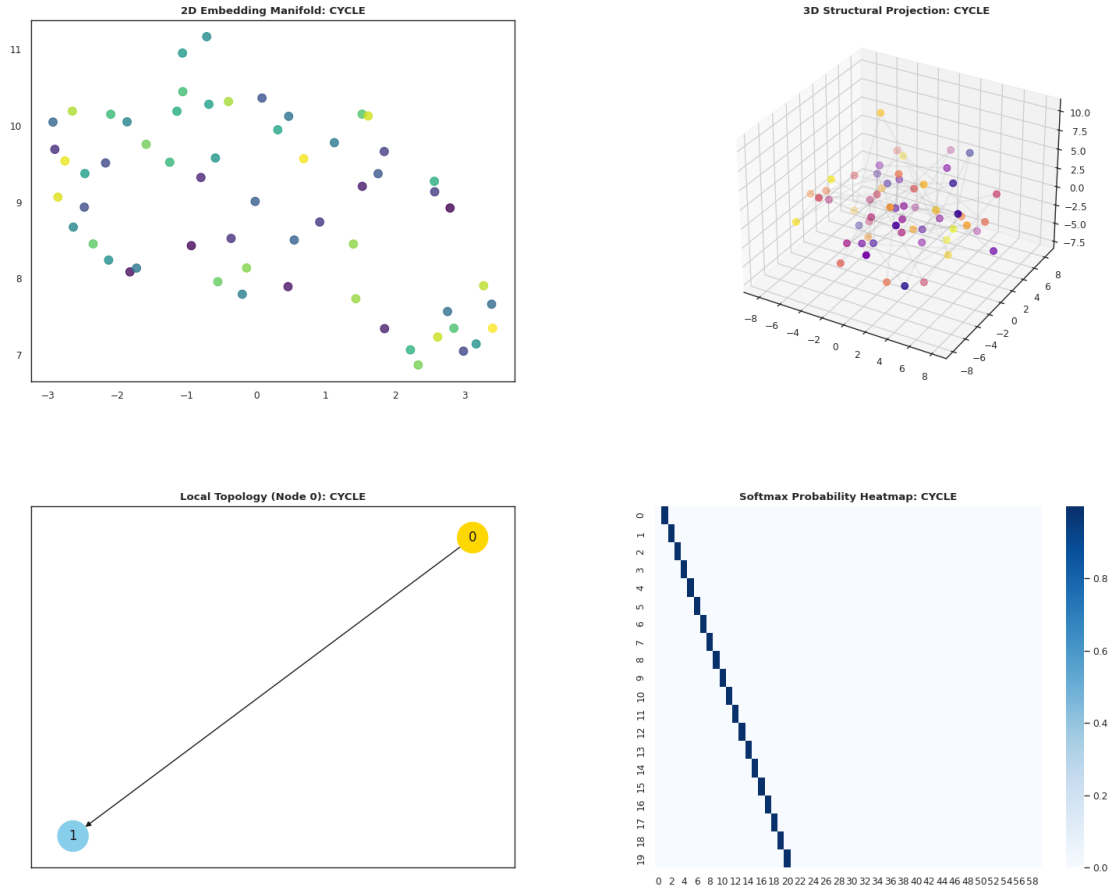

Figure 3: **Cycle graph: symmetry-induced non-geometric memorization.** Despite perfect memorization, embeddings do not exhibit a coherent global metric structure.

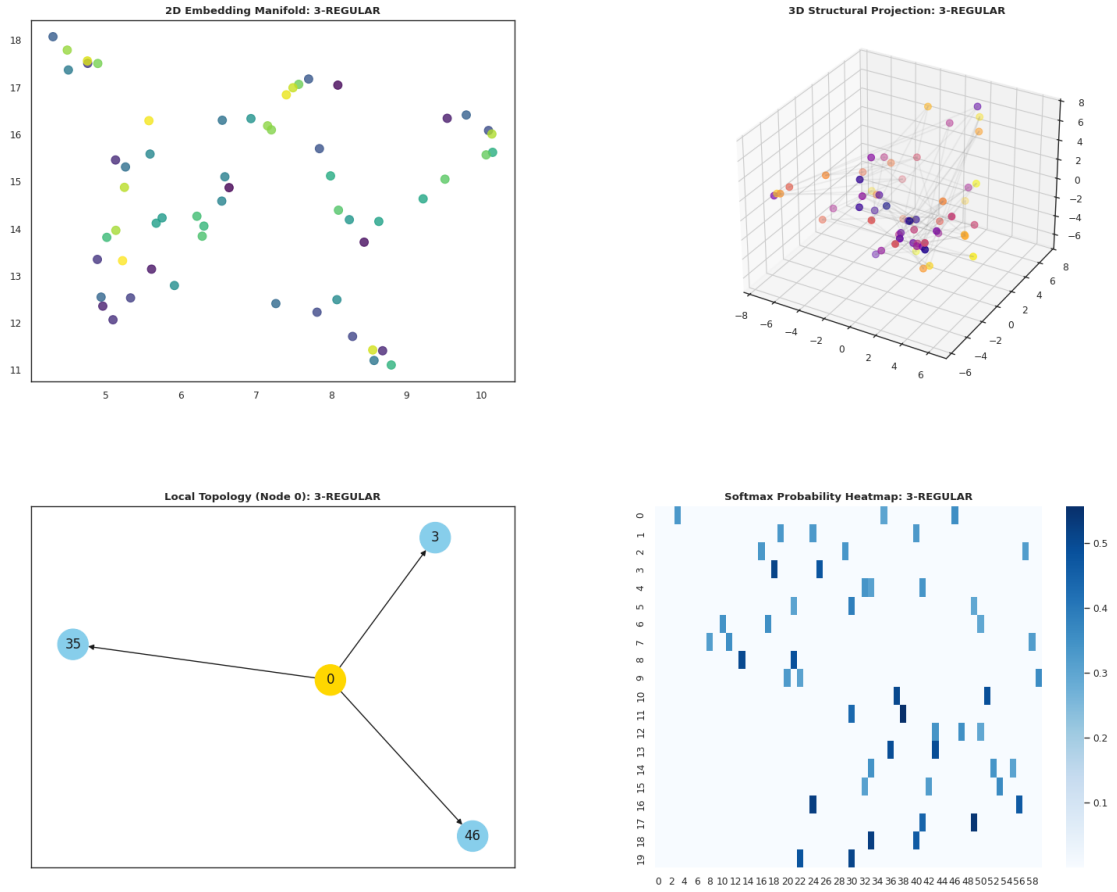

Figure 4: **3-Regular graph: partial asymmetry and fragile geometry.** Local neighborhood coherence emerges without a stable global metric, consistent with low-moderate rank correlation and elevated distortion.

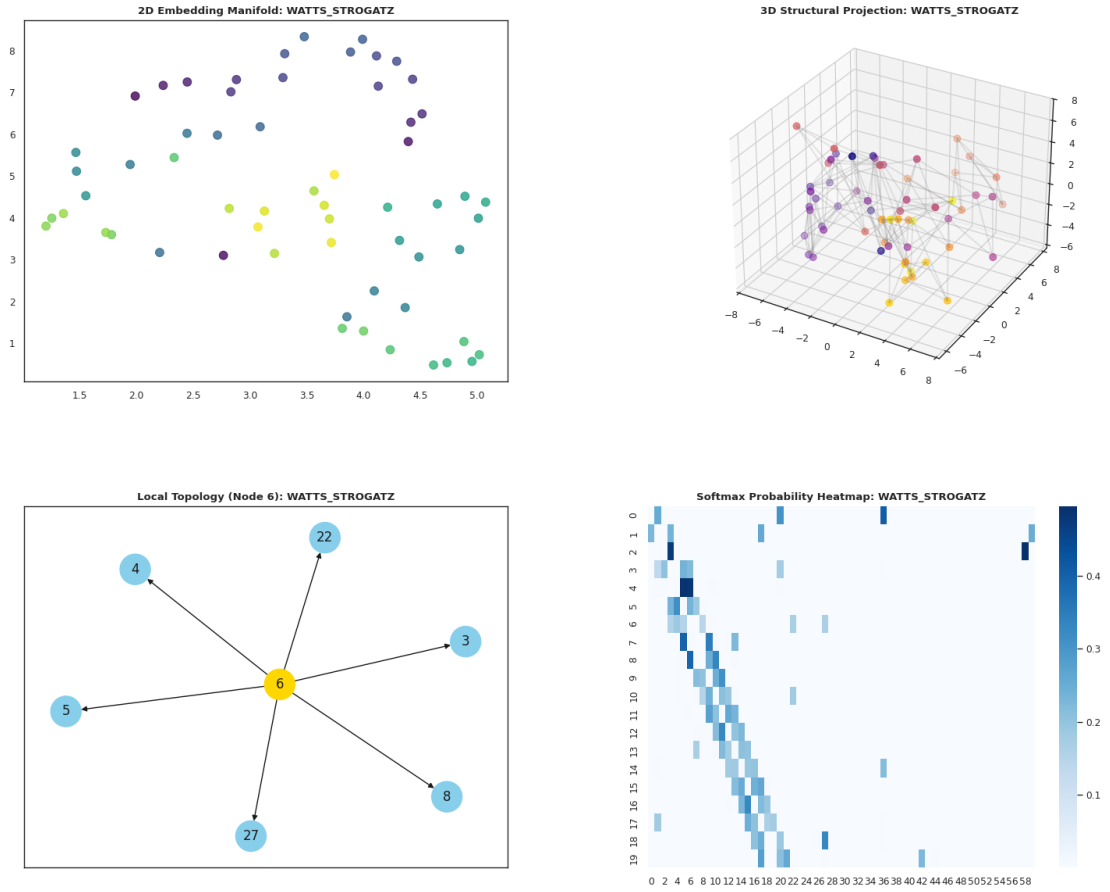

Figure 5: **Watts–Strogatz graph: emergent geometric organization.** Embeddings form a smooth low-dimensional structure that preserves graph distances, consistent with high rank correlation and controlled distortion.

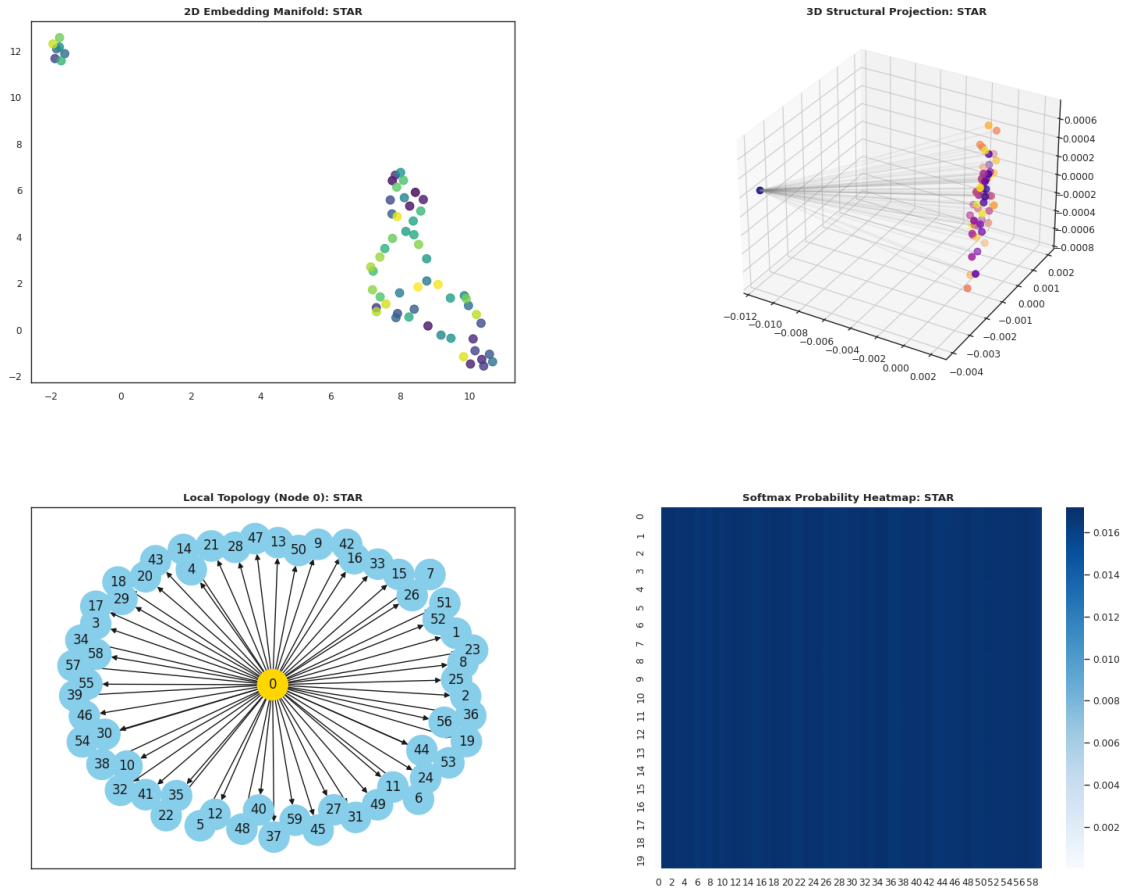

Figure 6: **Star graph: topology-determined geometric structure.** The hub–spoke topology yields a degenerate but task-sufficient geometric separation. Rank-based probes are undefined due to distance collapse, and the observed structure reflects topological degeneracy rather than recoverable metric organization.