What Constrains Adaptation After Pretraining? Generalization and Specialization Under Inherited Data Manifolds

Anonymous authors
Paper under double-blind review

Abstract

Large language models are often adapted to new tasks through supervised fine tuning. In deployment, however, their generalization can be unreliable and hard to anticipate. We examine whether such failures arise from limitations in optimization and supervision, or from geometric constraints inherited from pretraining, noting that data organization in representation space is rarely treated as an explicit control variable. Using controlled sampling from a large text distribution drawn from the web, we treat training samples as structured populations in representation space. We then compare data drawn from central and peripheral regions of the inherited manifold under identical architectures and training procedures. We find that data location in representation space strongly constrains what can be learned, frequently necessitating specialization across both general and domain-specific settings. Models trained on data drawn from peripheral or highly overlapping regions tend to generalize poorly, even when the training setup is otherwise unchanged. This pattern points to the need for principled specialization to meet practical demands on reliability, efficiency, and deployment.

1 Introduction

Applied systems must choose between reuse across settings and specialization for specific needs Caruana (1997). The difficulty becomes clear when shared representations are asked to support different and often conflicting requirements. Modern language models offer a particularly clear case. A single pretrained backbone is often reused across tasks and domains, with performance and reliability differing substantially across settings Raffel et al. (2020).

Modern language models derive their capabilities from vast training corpora, yet research attention has focused disproportionately on models rather than on data itself Bommasani et al. (2021). Over the past few years, the dominant training pipeline has largely stabilized. Pretraining on large collections of text drawn from the web Peters et al. (2018) is followed by supervised fine tuning Wei et al. (2022), often supplemented by distillation Bucilua et al. (2006); Hinton et al. (2015) or alignment procedures Ouyang et al. (2022). These stages are now well understood and widely adopted across model families.

Model behavior is often interpreted through objectives and optimization, while the organization of the training data remains largely unexamined. Training data are then processed as collections of samples, without regard to their arrangement in representation space. Correspondingly, a large body of work has framed progress in terms of model scaling Kaplan et al. (2020); Hoffmann et al. (2022), objective design van den Oord et al. (2018), and optimization algorithms Kingma & Ba (2015); Loshchilov & Hutter (2019). While the geometric organization of representations has received comparatively less explicit attention Rogers et al. (2020).

Related work has shown that learned representations have internal structure, and that this structure constrains how models behave Bronstein et al. (2017). Earlier work on distributed representations shows that

behavior depends on how internal states are organized, not on isolated symbolic rules Rumelhart & McClelland (1987). From this perspective, controllability and generalization in language models are best examined through the geometry of representation space, rather than solely through coarse behavioral metrics Elhage et al. (2021).

In this work, we treat training data as a structured population embedded in representation space. We examine how learning dynamics after pretraining are constrained by the geometry of the inherited data manifold.

We observe the same pattern across model sizes and across both fine tuning and distillation. Training that emphasizes boundary regions leads to lasting differences in generalization. These results indicate that data manifold geometry forms a distinct axis of control that limits what adaptation after pretraining can reliably change.

2 Related Work

Pretraining: Foundation of Language Models

Pretraining establishes the representation space on which later adaptation operates. Through large scale exposure to diverse text corpora, transformer models learn representations that capture broad linguistic structure Vaswani et al. (2017); Tennenholtz et al. (2024). This process organizes data into a structured representation space, beyond improvements in surface level task performance.

Prior work in representation learning shows that learned representations reflect both data structure and training objectives, rather than forming a uniform embedding space Bengio et al. (2013). Studies of neural population dynamics further suggest that high dimensional systems evolve along lower dimensional manifolds, which constrain subsequent learning dynamics Sussillo & Barak (2013); Cunningham & Yu (2014). Consistent with this view, adaptation after pretraining primarily redistributes weight within existing representations, while the overall geometric structure remains largely unchanged Raghu et al. (2017).

Beyond Pretraining: Adaptation and Its Constraints

Adaptation after pretraining modifies model behavior while operating on representations that are already established. Supervised fine tuning incorporates task supervision directly, whereas distillation shapes the student through guidance from a teacher model Gou et al. (2021).

Distillation is commonly applied by aligning output distributions Sun et al. (2019), intermediate representations Jiao et al. (2020); Wang et al. (2020), or attention patterns Tian et al. (2020). Across these approaches, adaptation does not introduce fundamentally new representational structure Mirzadeh et al. (2020); Hao et al. (2021). From a geometric perspective, adaptation after pretraining redistributes density within an inherited representation space, while the large scale structure remains unchanged Bengio et al. (2013).

Data Geometry in Representation Space

Training pipelines rarely make explicit how training data are situated in representation space. Contemporary work on large language models has focused on model architecture and training objectives Xiao & Zhu (2025). Scaling behavior is often treated as the principal catalyst of progress. In this framing, data are handled as collections of samples rather than as structured populations in representation space Bommasani et al. (2021).

Even when data are treated as a first class concern, they are typically handled through filtering, sampling, or efficiency driven choices, rather than through explicit control of representation geometry Wang et al. (2023); Albalak et al. (2024). Empirical studies of pretraining data Longpre et al. (2024) composition similarly emphasize domain coverage, data age, and quality, while leaving the geometric organization of data in representation space implicit.

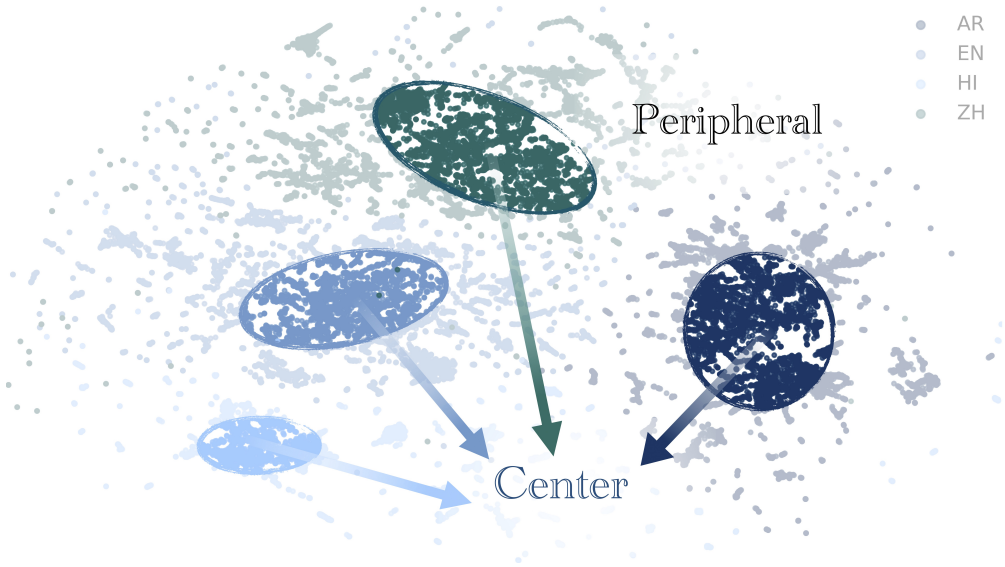

Figure 1: **Schematic illustration of center and peripheral sampling in representation space.** Training samples are selected either from high density core regions or from low density boundary regions of the inherited data manifold. The schematic highlights the geometric distinction underlying the quantile based partition described in Appendix A.4. The figure is illustrative and not drawn to scale.

A substantial body of prior work has shown that learned representations reside on structured, non Euclidean manifolds rather than forming uniform clouds of points Bronstein et al. (2017).

Classical manifold learning frameworks highlight that data geometry constrains the scope of downstream transformations Roweis & Saul (2000); Tenenbaum et al. (2000); Belkin & Niyogi (2003). Similar constraints have been observed in recent analyses of embedding organization and mechanistic structure in large language models Nanda et al. (2023).

Adaptation after pretraining operates under an inherited geometric structure. Training data drawn from different regions of the manifold give rise to distinct downstream behavior, even when training settings are held fixed. Fig. 1 illustrates this distinction by contrasting sampling from high density central regions of the data manifold with sampling from low density peripheral regions.

3 Experimental Design

We examine adaptation that follows pretraining under fixed training conditions. Variation reflects differences in geometric support within representation space.

Throughout our experimental setup, we explicitly control: *(i)* the model architecture and pretrained initialization, *(ii)* training data size, sequence processing, and supervision format, *(iii)* the adaptation operator (supervised fine tuning, LoRA, or distillation) together with its hyperparameters, and *(iv)* the evaluation protocol and representation extraction procedure defined in Appendix A. The sole independent variable is the geometric configuration of the training data, operationalized through neighborhood statistics computed in a fixed reference embedding space. Observed differences reflect variation in geometric support in representation space. Optimization, supervision, and model capacity are held fixed. In this sense, data geometry functions as an explicit control variable rather than an implicit byproduct of training.

3.1 Models and Datasets

Geometric analyses rely on representations extracted from the final layer of a pretrained decoder only language model f_θ. Unless otherwise noted, a sequence x is represented by the hidden state at the final non padding token, denoted as $z(x) = h_L(x)$.

The main study uses a pretrained model with 1B parameters, with selected analyses replicated using a 3B parameter model under matched training settings. We include three common post-pretraining adaptation operators. They are used as controlled instantiations of adaptation, rather than as objects of algorithmic comparison.

Supervised fine tuning (SFT). SFT updates all model parameters by minimizing the prediction loss for the next token on supervised training sequences. All SFT runs use the same data budget, input formatting, and optimization schedule across geometric conditions.

Low-Rank Adaptation (LoRA). LoRA limits parameter updates to low rank adapters applied to selected modules, while backbone weights remain fixed. Geometric effects are evaluated under a reduced number of trainable parameters.

Across operators, the emphasis is not on maximizing absolute task performance. Instead, we examine whether geometric effects remain consistent under common adaptation mechanisms applied after pretraining.

Dataset Geometry and Conditioned Sampling Training data are drawn from a large public text distribution. Rather than treating the dataset as an unordered collection of samples, we treat candidate samples as a structured population embedded in a fixed reference embedding space. This reference space is used exclusively for dataset organization and sampling; all adaptation analyses are performed on decoder-only representations.

Reference embedding space. Candidate samples are embedded with a fixed sentence embedding model (BGE) to define a stable reference geometry for the dataset. Let $\mathcal{Z}_{\text{ref}} = \{u_i\}_{i=1}^{N}$ denote the resulting reference embeddings. The fixed encoder yields a consistent space for density estimation and sampling, independent of the representations of f_θ that evolve during adaptation.

Density proxy and quantile sampling. Local density is approximated using the mean k-nearest neighbor distance:

$$\rho_k(u_i) = \frac{1}{k} \sum_{u_j \in \mathcal{N}_k(u_i)} \|u_i - u_j\|_2, \tag{1}$$

where $\mathcal{N}_k(u_i)$ denotes the set of k nearest neighbors under Euclidean distance. Smaller $\rho_k(u_i)$ corresponds to denser (central) regions of the dataset manifold, while larger values indicate peripheral regions.

Paired datasets are constructed under identical budgets using fixed quantile thresholds.

$$\begin{aligned} \mathcal{D}_{\text{center}} &= \{x_i : \rho_k(u_i) \le q_{\text{low}}\}, \\ \mathcal{D}_{\text{periph}} &= \{x_i : \rho_k(u_i) \ge q_{\text{high}}\}. \end{aligned} \tag{2}$$

The quantiles $(q_{\text{low}}, q_{\text{high}})$ are fixed across models, operators, and checkpoints. Samples are drawn from $\mathcal{D}_{\text{center}}$ and $\mathcal{D}_{\text{periph}}$. The two sets are matched in size, while formatting and supervision remain unchanged. This design isolates geometric support as the only source of variation.

Rationale for a separate reference space. A separate reference embedding space is used to define dataset level geometry and sampling. This choice decouples data selection from representations that change during adaptation. All geometric analyses reported in this work are conducted in the decoder only representation space of f_θ.

Role of visualization UMAP visualizations are included solely to provide qualitative context. All conclusions are based on quantitative statistics that are invariant to scale, which are reported in later sections.

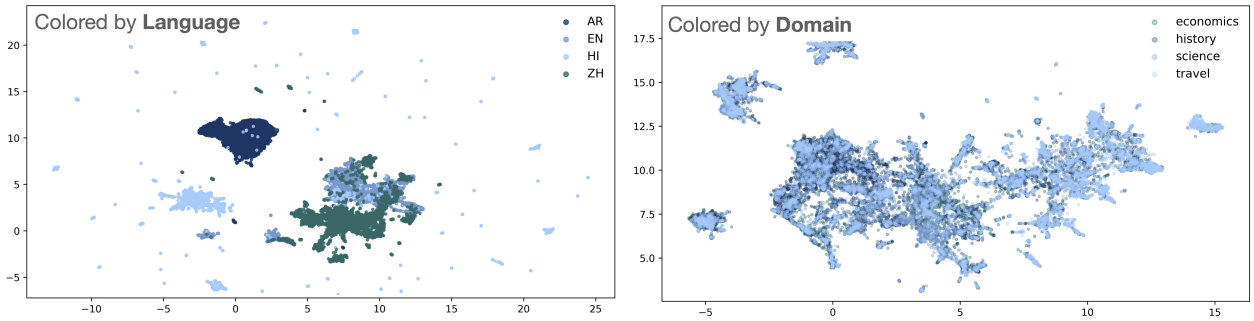

Figure 2: **Dataset geometry under different annotation axes. (a)** UMAP of dataset embeddings colored by language, showing well separated clusters. **(b)** The same embeddings colored by coarse domain labels, which exhibit substantial overlap. Embeddings are computed using a fixed encoder-based sentence embedding model.

3.2 Language and Domain Structure in Representation Space

We focus on two sources of annotation, **language** and **domain**. As shown in Fig. 2, language labels correspond to well separated regions in representation space, whereas coarse domain labels occupy substantially overlapping regions. This contrast motivates the hypothesis examined in this section. Adaptation after pretraining redistributes density within geometrically separable submanifolds, and is more constrained when representations occupy entangled regions.

3.3 Evaluation Metrics

The evaluation considers three complementary aspects: *utility*, *generalization*, and *representation geometry*. All metrics are computed on held out data that are disjoint from the geometry conditioned training subsets.

Utility. Task performance is summarized by token level negative log likelihood (NLL), reported on in distribution evaluation splits. Decoding and batching settings are kept the same across all conditions.

Generalization and specialization. To assess generalization outside the training support, we evaluate models trained on $\mathcal{D}_{\text{center}}$ or $\mathcal{D}_{\text{periph}}$ using complementary data regions. Specialization is quantified by the performance gap between data within the training support and data outside it:

$$\Delta_{\text{gen}} = \text{NLL}(\text{outside support}) - \text{NLL}(\text{within support}).$$

Complementary regions are defined using the same quantile thresholds in a fixed reference embedding space. This yields regions with matched support size that occupy disjoint locations in representation space.

Geometric evaluation. Representation geometry is assessed using statistics computed from final layer embeddings. Global measures include $\mathbb{E}[\|z\|_2]$, $\mathbb{E}[d_2]$, $\mathbb{E}[d_{\cos}]$, and $\mathbb{E}[d_2^{\text{dir}}]$. Local structure is summarized through k nearest neighbor statistics, including the mean, standard deviation, coefficient of variation, and upper tail quantiles of ρ_k. All geometric quantities are estimated using a fixed subsampling budget of M random pairs and a shared random seed across checkpoints.

Baselines and Comparisons We compare geometry-conditioned sampling against the following baselines under identical training budgets: **(i) Random sampling**; **(ii) Quality-filtered sampling**; **(iii) Similarity-based sampling without density conditioning**; and **(iv) Curriculum variants**. All baselines share the same formatting, sequence processing, and optimization settings.

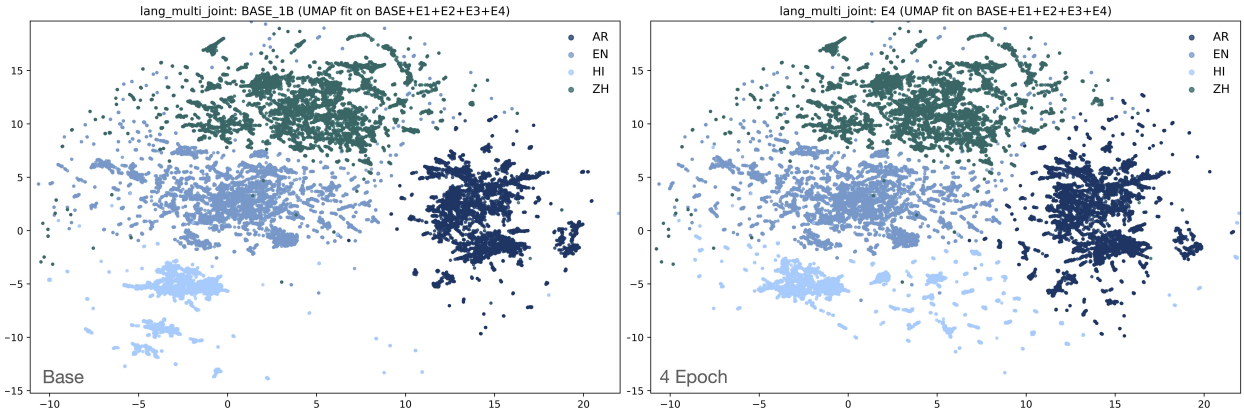

Figure 3: **Early geometric effects of language-conditioned fine tuning.** Joint fit UMAP projections of final layer embeddings from a 1B parameter model, shown before fine tuning (baseline) and after four epochs of supervised fine tuning (E4). Fine tuning maintains the separation between language clusters while smoothing density within each cluster, without inducing cross language collapse.

3.4 Implementation Details

All experiments are repeated *[R]* times with different random seeds, and we report mean $\pm$ standard deviation. Unless stated otherwise, we use *[model name]* at scales 1B and 3B. Hyperparameters for SFT, LoRA, and distillation follow standard practice and are held constant across geometric conditions.

Representation Extraction for Geometry Analysis Representations are taken from the final layer at the last non padding token for each input sequence x and each training checkpoint, yielding the embedding $z(x) = h_L(x)$. Unless otherwise noted, analyses are performed on raw embeddings. Additional statistics are computed on L_2 normalized embeddings to separate scale sensitive effects from directional structure.

4 Experiments and Evaluation

To disentangle geometric structure at the level of the dataset from adaptation dynamics specific to a given model, we distinguish between representations used for data analysis and those used for training analysis.

4.1 Early Geometric Effects of Post-Pretraining Adaptation

We first examine how supervised fine tuning alters the geometry of representation space in a multilingual setting. Fig. 3 presents joint fit Uniform Manifold Approximation and Projection (UMAP) visualizations McInnes et al. (2018) of final layer embeddings extracted from a 1B parameter model before fine tuning

Epoch	Mean norm	Norm std	Mean L2 dist	Mean cosine dist	Mean dir. dist
Base	61.826	4.942	71.167	0.716	1.146
E1	59.356	6.103	69.819	0.740	1.169
E2	55.628	5.388	66.116	0.750	1.182
E3	53.550	5.139	64.028	0.757	1.190
E4	52.929	5.095	63.378	0.759	1.192

Table 1: **Changes in global representation geometry across fine tuning epochs (language-conditioned, 1B model).** Mean embedding norm, global Euclidean distance, cosine distance, and direction only Euclidean distance computed from last layer embeddings. Values are averaged over randomly sampled pairs under identical evaluation settings.

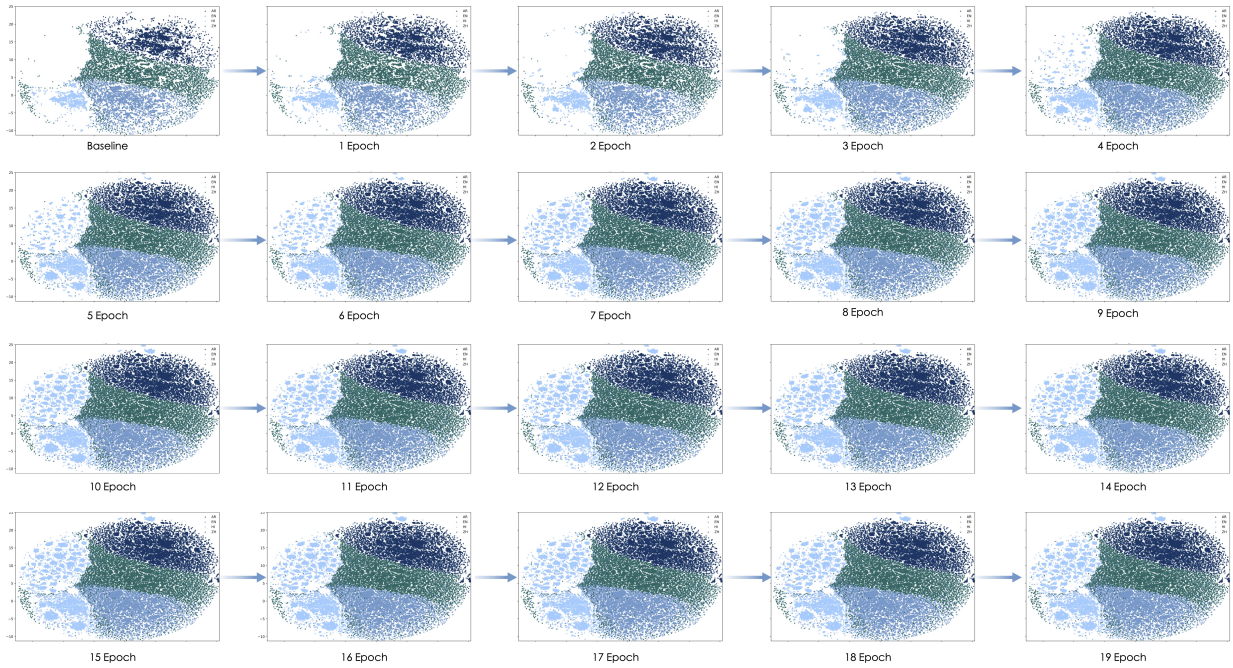

Figure 4: **Temporal evolution of representation geometry under extended fine tuning.** Joint fit UMAP projections of final layer embeddings from the baseline model and checkpoints from epochs 1 through 19, shown using a shared projection and fixed axis limits. The global organization of representations emerges within the early epochs, with subsequent training primarily redistributing density inside existing regions.

(baseline) and after four epochs of language conditioned supervised fine tuning (E4). Both panels rely on the same UMAP projection with fixed hyperparameters and axis limits. As a result, differences reflect geometric changes rather than visualization effects.

UMAP projections reveal stable relative positions of language clusters after supervised fine tuning, alongside redistribution of density within clusters. We assess global distance statistics across training epochs to clarify whether the observed behavior reflects representational collapse or a contraction in scale (Table 1).

Across training epochs, embeddings contract in scale. The mean norm drops from 61.826 at baseline to 52.929 after four epochs of fine tuning, accompanied by a parallel decrease in average Euclidean pairwise distance from 71.167 to 63.378.

This contraction does not reflect a loss of directional structure. Instead, cosine based global distances increase from 0.716 to 0.759, and direction only Euclidean distances computed on normalized embeddings rise from 1.146 to 1.192. Together, these trends indicate that angular separation between representations is preserved, and in fact becomes slightly more pronounced.

Taken together, the joint fit UMAP visualizations and global distance statistics indicate scale contraction with preserved, and slightly increased, angular separation. Under language conditioning, fine tuning mainly redistributes density within existing language regions, while their overall separation remains intact.

These early observations motivate a broader question addressed next. A natural next question is how far post-pretraining adaptation can go, and whether it mainly redistributes density or can also reshape the underlying manifold. We therefore turn to axis dependent effects and quantify how manifold redistribution differs across annotation dimensions.

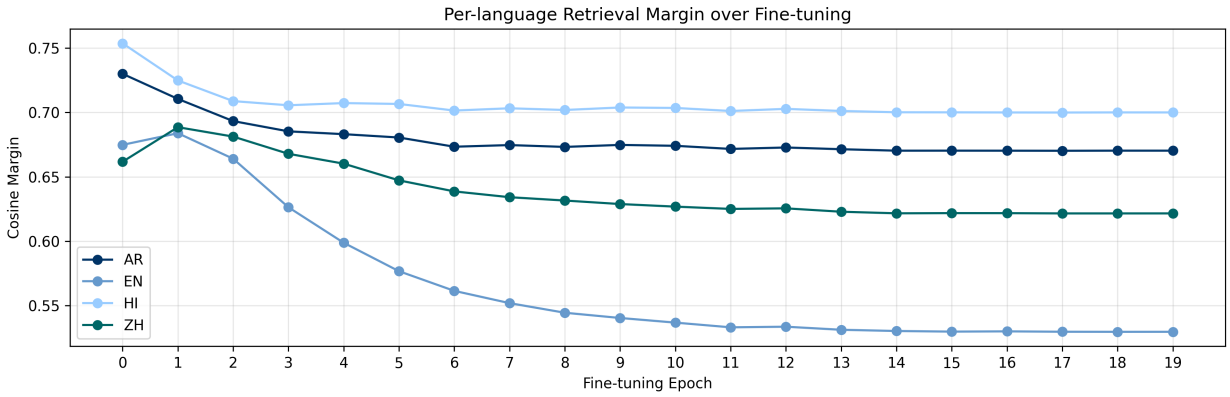

Figure 5: **Per language cosine margin under extended fine tuning.** Mean centroid based cosine margin for English(EN) , Chinese(ZH), Hindi(HI), and Arabic(AR) from the baseline through E19. Top one language retrieval remains perfect across checkpoints, while the margin decreases and stabilizes after approximately E13.

4.2 Stability under Extended Fine Tuning

The geometric patterns observed during early adaptation persist under extended training. Supervised fine tuning is continued for up to twenty epochs under the same language conditioned setting, allowing changes observed after a small number of epochs to be examined over a longer training horizon.

Fig. 4 presents joint fit UMAP projections of final layer embeddings from the same 1B model, covering checkpoints from the baseline through E19. A shared projection and fixed axis limits are used across checkpoints.

Separability is quantified using a centroid based cosine margin for language identification. Top one language retrieval remains perfect across all checkpoints, indicating that language clusters do not collapse. The mean margin decreases monotonically from 0.705 at baseline to 0.630 at E19 and stabilizes after approximately E13. This trend reflects a gradual reduction in inter language separation while preserving overall cluster separability. Detailed statistics for individual languages are reported in Appendix B.

Across training, language clusters maintain their relative positions. Density shifts occur within individual language regions, while the global organization of the representation manifold remains stable. Beyond the early epochs, geometric changes diminish in magnitude, indicating convergence toward a stable configuration rather than a delayed collapse.

These observations support the interpretation that adaptation after pretraining primarily refines an inherited representation manifold. Extending fine tuning over many epochs does not alter the geometric constraints established during pretraining, but instead reinforces structure that is already present.

4.3 Scale Dependence and the Role of Manifold Separability

We apply the same joint fit analysis to a 3B parameter model under identical training conditions to assess whether the observed geometric effects persist with increased capacity. Figure 6 compares language conditioned and domain conditioned settings before and after supervised fine tuning.

The qualitative asymmetry observed at 1B remains evident at larger scale. When embeddings are organized by language, the regions remain clearly separated. This organization can already be seen in the pretrained model. Supervised fine tuning mainly adjusts how representations are distributed within each region. In contrast, domain conditioned embeddings continue to occupy a shared and highly overlapping region, even after fine tuning.

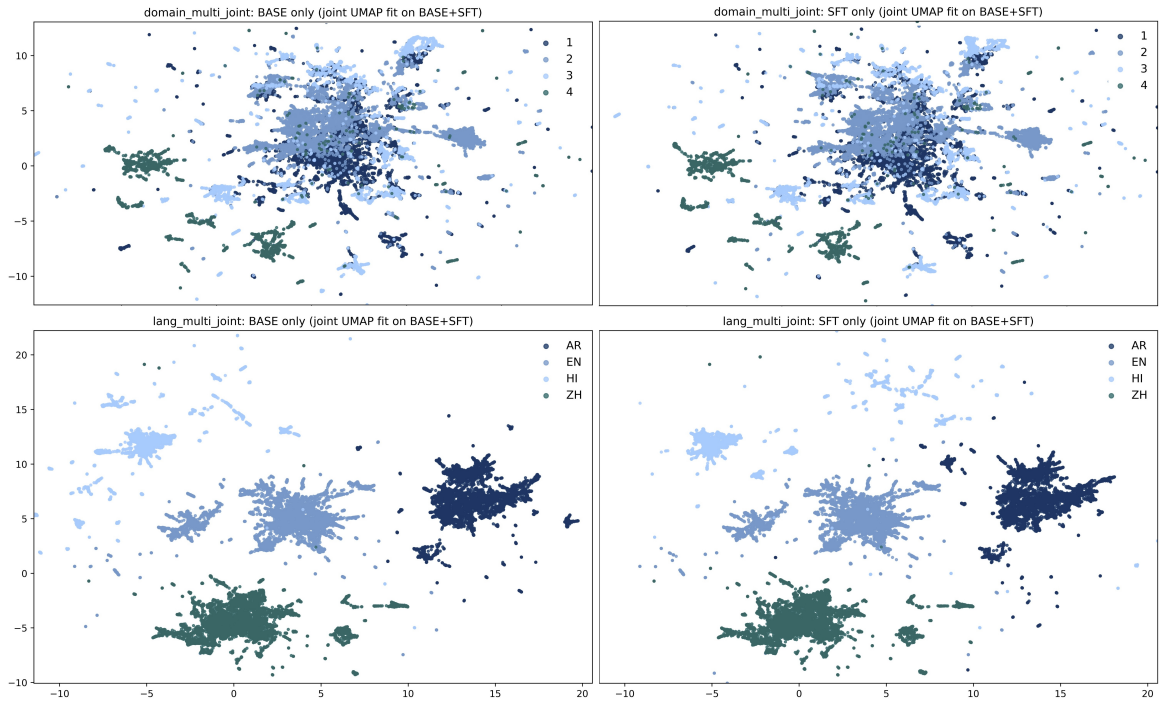

Figure 6: **Geometric effects of supervised fine tuning across model scale.** Joint fit UMAP projections of final layer embeddings from a 3B model, shown for language conditioned and domain conditioned settings before and after supervised fine tuning. Language based organization yields distinct regions with density redistribution under training, whereas domain based organization remains largely overlapping with limited geometric change.

These observations indicate that the effectiveness of post pretraining adaptation depends critically on the existing separability of the data manifold. Increasing model capacity alone does not enable fine tuning to reorganize intrinsically overlapping regions of representation space. Instead, adaptation continues to operate within the geometric structure established during pretraining. The same pattern observed at 1B also appears at larger scale, indicating a geometric constraint that remains consistent across model sizes.

4.4 Quantifying Manifold Redistribution under Post-Pretraining Adaptation

The qualitative asymmetries observed in Fig. 2 suggest that post-pretraining learning does not act uniformly across representation space. To quantify this effect, we compute global and local geometric statistics from last layer embeddings under different training regimes.

These patterns are reflected in the statistics reported in Table 2. Across all settings, the global measures follow a consistent pattern. Average pairwise distances decrease under supervised fine tuning and distillation in both domain conditioned and language conditioned settings. This decrease appears gradually over training and is consistent across model sizes.

Local structure, however, behaves very differently depending on the annotation axis. Under domain conditioned sampling, neighborhoods become progressively more uniform. Both dispersion and tail distances decrease across models and training regimes, as local geometry smooths within an already overlapping space.

Language conditioned settings show the opposite behavior. Adding languages leads to a looser neighborhood structure, while large distances continue to appear. The same pattern holds across model sizes and adaptation methods, with representations remaining separated by language. Adaptation modifies density within each region rather than pulling them into a single collective environment.

Table 2: **Representation geometry under domain-conditioned adaptation.** Global and local statistics computed from final-layer embeddings under domain-based sampling.

Domain	Scale	Training	NLL	global_intra	std	knn_mean	knn_std	knn_cv	knn_q90
x1	1B	SFT	2.050	42.823	6.581	27.643	6.662	0.241	36.101
	3B	–	2.085	37.851	5.419	25.738	5.988	0.233	32.692
	3B	LoRA	1.923	38.111	5.593	25.356	6.155	0.243	32.354
	3B	SFT	1.874	37.839	5.529	25.158	6.152	0.245	32.248
x2	1B	SFT	2.188	47.921	7.013	34.469	6.997	0.203	43.865
	3B	–	2.187	42.302	5.860	31.589	6.079	0.192	39.745
	3B	LoRA	2.036	42.885	5.973	31.679	6.272	0.198	39.826
	3B	SFT	1.997	42.429	5.946	31.294	6.275	0.201	39.594
x3	1B	SFT	2.139	48.796	6.342	35.157	6.441	0.183	43.469
	3B	–	2.148	42.730	5.148	32.160	5.569	0.173	39.324
	3B	LoRA	1.994	43.328	5.259	32.228	5.799	0.180	39.631
	3B	SFT	1.948	42.850	5.198	31.873	5.753	0.180	39.059
x4	1B	SFT	2.091	48.682	6.271	35.030	6.350	0.181	42.335
	3B	–	2.091	42.512	5.010	31.970	5.414	0.169	38.639
	3B	LoRA	1.940	43.177	5.171	32.066	5.650	0.176	38.840
	3B	SFT	1.892	42.703	5.093	31.721	5.638	0.178	38.331

Post-pretraining adaptation reshapes representation geometry differently across annotation axes. Global measures reflect a normalization of scale, whereas local statistics indicate how learning is organized in representation space. In domain conditioned settings, learning unfolds within a single overlapping space, while in language conditioned settings it spans multiple separated regions. Fine tuning smooths structure in domain conditioned settings. Language specific organization, by contrast, shows little change.

5 Discussion

Post-pretraining adaptation does not rearrange representation space wholesale. Rather than restructuring representation space, adaptation acts on structure that is already in place, shifting where representations concentrate. This behavior has direct implications for specialization. Alignment between an annotation axis and inherited geometry allows representations to separate during training. Without such alignment, additional supervision instead concentrates representations within the same region. Throughout the analysis, we use the center–periphery partition defined in Appendix A.

Post-Pretraining Adaptation as Geometric Redistribution

This pattern appears consistently across model sizes and training regimes. Adaptation leaves the large-scale organization of representation space largely intact. Instead, learning primarily redistributes density within an inherited manifold. Joint-fit projections show stable relative positions of major regions, while quantitative statistics reveal contraction in scale without collapse of directional structure. These patterns indicate that fine tuning and distillation operate within geometric constraints established during pretraining.

Geometric Alignment and the Limits of Specialization

Language and domain conditioning differ in ways that track their alignment with representation geometry. Adaptation is most effective when the annotation axis aligns with structure already present in representation space. Language provides such an axis, as it gives rise to coherent regions that fine tuning can refine without altering the global organization.

Coarse topical domains overlap extensively and span several latent dimensions, such as style, intent, and register. If labels do not align with separable regions of the manifold, post-pretraining learning unfolds within a shared and entangled geometry. Specialization in this regime is therefore weak and largely uniform. Instead, the limits of fine tuning appear to lie elsewhere. They reflect a mismatch between supervision labels and the structure already present in representation space.

Geometric Constraints as a Design Principle

The importance of geometric constraints is increasingly recognized not only at the data level but also at the architectural level. Recent large scale architecture work has shown that explicitly constraining residual connections to lie on well defined manifolds is essential for stable and scalable training Xie et al. (2025). Although operating at different levels, these architectural results are conceptually aligned with our findings. Both suggest that unconstrained transformations of representation space are neither stable nor effective at scale.

The results point to data manifold geometry as its own axis of control, alongside model capacity and objective design. Understanding where training data reside within representation space is therefore essential for interpreting both the capabilities and the limits of post-pretraining adaptation.

Implications for Domain Focused and Capacity Constrained Models

The findings point to a practical implication for improving reliability in models with limited capacity or a focused deployment domain. When capacity is constrained, generalization degrades as a model is required to span increasingly wide and heterogeneous regions of the inherited representation manifold. In such regimes, performance degradation reflects geometric overextension rather than insufficient optimization or supervision.

Reliability improves when both training and deployment are confined to geometrically coherent regions of the pretrained manifold. Selecting data that occupy dense and well supported subregions aligned with a specific domain allows adaptation to operate within a clean and internally consistent geometric structure. Regions that are strongly supported by pretraining preserve angular structure with reduced interference from overlapping or peripheral representations.

This notion of maintaining a clean model within a given domain differs from conventional data cleaning or deduplication. The limiting factor is geometric rather than lexical. Stable behavior emerges when adaptation remains within the boundaries of the inherited manifold, rather than relying on surface level similarity. Under this constraint, specialization follows naturally, and compact models exhibit more reliable behavior when operating within regions of representation space that are strongly supported by pretraining.

Boundary of the Analysis

The analysis focuses on representations extracted from the final layer and does not trace geometric variation across depth. Prior work indicates that later layers encode task relevant structure, and the present study examines how adaptation redistributes geometry within this representational regime.

The conclusions are specific to adaptation after pretraining under fixed data support. The analysis does not address scenarios in which substantially different data distributions or objectives induce new representational structure. Within the settings considered here, including supervised fine tuning and distillation, learning proceeds under geometric constraints inherited from pretraining.

Visualization serves to provide qualitative context for local structure. All reported effects are supported by scale invariant quantitative measures, and the same geometric patterns are observed across both forms of analysis. This consistency indicates that the findings are not driven by projection artifacts.

In multi domain deployment under constrained capacity, a single model must span heterogeneous regions of representation space. Under such conditions, specialization follows from the organization of the inherited representation geometry.

Model Scale and Scope

The study considers models at two representative scales, with 1B and 3B parameters. The objective is not to establish state of the art performance, but to isolate geometric constraints that govern adaptation under controlled conditions.

Smaller and mid scale models provide a clearer view of representation geometry, making changes in manifold structure, density redistribution, and separability easier to observe. Across both model scales, and across supervised fine tuning and distillation, the same qualitative behaviors are consistently observed. This consistency indicates that the reported effects reflect structural properties of representation geometry rather than artifacts of limited capacity.

While larger models may admit additional degrees of freedom, increased scale alone does not remove geometric constraints imposed by inherited data organization. The central findings are therefore expected to extend beyond the specific scales examined here, particularly in settings that require compact or specialized models due to reliability, efficiency, or deployment considerations.

Acknowledgments

The author thanks colleagues for many helpful discussions and comments throughout this work.

References

Alon Albalak, Yanai Elazar, Sang Michael Xie, Shayne Longpre, Nathan Lambert, Xinyi Wang, Niklas Muennighoff, Bairu Hou, Liangming Pan, Haewon Jeong, Colin Raffel, Shiyu Chang, Tatsunori Hashimoto, and William Yang Wang. A survey on data selection for language models. *arXiv preprint arXiv:2402.16827*, 2024. doi: 10.48550/arXiv.2402.16827. URL `https://arxiv.org/abs/2402.16827`. cs.CL.

Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. In *Neural Information Processing Systems*, pp. 585–591, 2003.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013. doi: 10.1109/TPAMI.2013.50.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, and William Wang. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021. Version 3, revised July 2022.

Michael M. Bronstein, Joan Bruna, Yann LeCun, Arthur Szlam, and Pierre Vandergheynst. Geometric deep learning: Going beyond euclidean data. *IEEE Signal Processing Magazine*, 34(4):18–42, 2017. doi: 10.1109/MSP.2017.2693418.

C. Bucilua, R. Caruana, and A. Niculescu-Mizil. Model compression. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 535–541, 2006. doi: 10.1145/1150402.1150464.

Rich Caruana. Multitask learning. *Machine Learning*, 28(1):41–75, 1997. doi: 10.1023/A:1007379606734.

John P. Cunningham and Byron M. Yu. Dimensionality reduction for large-scale neural recordings. *Nature Neuroscience*, 17(11):1500–1509, 2014.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Jackson Kernion, Kamal Ndousse, Liane Lovitt, Nova Tran-Johnson, Andy Mueller, Ben Landau, Esben Perez, Sam Schubert, Chris Hallacy, Amanda Askell, et al. A mathematical framework for transformer circuits. `https://transformer-circuits.pub/2021/framework/index.html`, 2021.

Jianping Gou, Baosheng Yu, Stephen J. Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6):1789–1819, 2021.

Zhiwei Hao, Jianyuan Guo, Ding Jia, Kai Han, Yehui Tang, Chao Zhang, Han Hu, and Yunhe Wang. Learning efficient vision transformers via fine-grained manifold distillation. *arXiv:2107.00374*, 2021. Version 4, last revised 2 Jun 2022.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv:1503.02531*, 2015.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. Tinybert: Distilling BERT for natural language understanding. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4163–4174, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.372.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations (ICLR)*, 2015.

Shayne Longpre, Gregory Yauney, Emily Reif, Katherine Lee, Adam Roberts, Barret Zoph, Denny Zhou, Jason Wei, Kevin Robinson, David Mimno, and Daphne Ippolito. A pretrainer's guide to training data: Measuring the effects of data age, domain coverage, quality, and toxicity. *arXiv preprint arXiv:2305.13169*, 2024. doi: 10.48550/arXiv.2305.13169. URL `https://arxiv.org/abs/2305.13169`. cs.CL.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

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

Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, and Hamid Ghasemzadeh. Improved knowledge distillation via teacher assistant. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 5191–5198, 2020.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. *arXiv preprint arXiv:2301.05217*, 2023. doi: 10.48550/arXiv.2301.05217. URL https://arxiv.org/abs/2301.05217.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*, 2022. doi: 10.48550/arXiv.2203.02155.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of NAACL-HLT*, pp. 2227–2237, 2018. doi: 10.18653/v1/N18-1202.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL http://jmlr.org/papers/v21/20-074.html.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 6078–6087, 2017.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866, 2020. doi: 10.1162/tacl_a_00349. URL https://aclanthology.org/2020.tacl-1.54/.

Sam T Roweis and Lawrence K Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

David E. Rumelhart and James L. McClelland. Chapter 3: Distributed representations. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Foundations*, pp. 75–109. MIT Press, Cambridge, MA, 1987.

Siqi Sun, Yu Cheng, Zhe Gan, and Jingjing Liu. Patient knowledge distillation for BERT model compression. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pp. 4314–4323, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1441.

David Sussillo and Omri Barak. Opening the black box: Low-dimensional dynamics in high-dimensional recurrent neural networks. *Neural Computation*, 25(3):626–649, 2013.

Joshua B Tenenbaum, Vin De Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

Guy Tennenholtz, Yinlam Chow, Chih-Wei Hsu, Jihwan Jeong, Lior Shani, Azamat Tulepbergenov, Deepak Ramachandran, Martin Mladenov, and Craig Boutilier. Demystifying embedding spaces using large language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024. doi: 10.48550/arXiv.2310.04475. Accepted to ICLR 2024.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. In *International Conference on Learning Representations (ICLR)*, 2020.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv:1807.03748*, 2018. doi: 10.48550/arXiv.1807.03748.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pp. 6000–6010, 2017.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *arXiv:2002.10957*, 2020. doi: 10.48550/arXiv.2002.10957.

Zige Wang, Wanjun Zhong, Yufei Wang, Qi Zhu, Fei Mi, Baojun Wang, Lifeng Shang, Xin Jiang, and Qun Liu. Data management for training large language models: A survey. *arXiv preprint arXiv:2312.01700*, 2023. doi: 10.48550/arXiv.2312.01700. URL `https://arxiv.org/abs/2312.01700`. cs.CL.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2022. doi: 10.48550/arXiv.2109.01652.

Tong Xiao and Jingbo Zhu. Foundations of large language models. *arXiv preprint arXiv:2501.09223*, 2025. doi: 10.48550/arXiv.2501.09223. URL `https://arxiv.org/abs/2501.09223`. cs.CL.

Zhenda Xie, Yixuan Wei, Huanqi Cao, Chenggang Zhao, Chengqi Deng, Jiashi Li, Damai Dai, Huazuo Gao, Jiang Chang, Liang Zhao, Shangyan Zhou, Zhean Xu, Zhengyan Zhang, Wangding Zeng, Shengding Hu, Yuqing Wang, Jingyang Yuan, Lean Wang, and Wenfeng Liang. mhc: Manifold-constrained hyper-connections. *arXiv preprint arXiv:2512.24880*, 2025.

A Preliminary Knowledge

This section introduces the representation and geometric quantities used throughout the paper. We describe how representations are extracted from a pretrained decoder only language model, the distance measures used to characterize geometry while separating scale from direction, the neighborhood statistics that summarize local structure, and the quantile partitions used to define center and periphery subsets. The definitions provided here are descriptive in nature and establish evaluation objects that are reused in later sections.

A.1 Sequences, Models, and Layer Representations

Consider a token sequence $x = (x_1, \ldots, x_T)$ of length T and a pretrained decoder only language model f_θ with parameters θ and L transformer layers. At each layer $\ell \in \{1, \ldots, L\}$, the model produces a hidden representation $h_\ell(x) \in \mathbb{R}^d$, where d denotes the hidden dimension.

Unless otherwise noted, each sequence x is represented by the hidden state at the final non padding token of the last layer,

$$z(x) \coloneqq h_L(x) \in \mathbb{R}^d. \tag{3}$$

This representation provides a consistent readout across variable length sequences under autoregressive processing and captures higher level semantic structure encoded in later layers. Using a single vector per sequence also simplifies subsequent geometric analysis by avoiding additional pooling choices.

For a dataset $\mathcal{D} = \{x_i\}_{i=1}^N$, the corresponding set of representations is denoted by

$$\mathcal{Z} \coloneqq \{z_i = z(x_i)\}_{i=1}^N \subset \mathbb{R}^d. \tag{4}$$

We refer to $\mathcal{Z}$ as an empirical data manifold induced by the pretrained model. This term denotes the observed distribution of samples in representation space, without assuming a parametric form, smoothness, or global manifold regularity beyond what is empirically supported. When comparing training checkpoints, $\mathcal{Z}^{(t)}$ denotes the representation set extracted from checkpoint t.

A.2 Distances, Scale, and Direction

Complementary distance measures are used to separate scale effects from directional structure. For two embeddings $z, z' \in \mathbb{R}^d$, Euclidean distance is

$$d_2(z, z') \coloneqq \|z - z'\|_2. \tag{5}$$

Euclidean distance depends on both magnitude and direction and therefore captures global scale contraction and norm changes.

To quantify directional relationships, cosine distance is

$$d_{\cos}(z, z') \coloneqq 1 - \frac{\langle z, z' \rangle}{\|z\|_2 \, \|z'\|_2}, \tag{6}$$

which is invariant to uniform rescaling and isolates differences in direction.

Directional geometry can also be summarized using an L_2 metric after normalization. Let $\tilde{z} \coloneqq z/\|z\|_2$. Then

$$d_2^{\mathrm{dir}}(z, z') \coloneqq \|\tilde{z} - \tilde{z}'\|_2. \tag{7}$$

This metric provides an equivalent view of angular separation up to a monotone transform, while remaining in Euclidean form. Unless otherwise noted, embeddings are analyzed in raw form so that norm contraction and global scale changes remain observable.

For dataset level summaries, global scale is tracked using the mean norm $\mathbb{E}[\|z\|_2]$ and its dispersion. Global geometry is summarized using expected pairwise distances $\mathbb{E}[d_2]$ and $\mathbb{E}[d_{\cos}]$. Because exact pairwise computation scales as $O(N^2)$, these expectations are estimated by random subsampling of pairs under a fixed sample size and a shared random seed across checkpoints and settings.

A.3 k Nearest Neighbor Neighborhoods and Local Density Proxies

Local structure is described using k nearest neighbor neighborhoods under Euclidean distance in $\mathbb{R}^d$. For each representation $z_i \in \mathcal{Z}$, the neighborhood $\mathcal{N}_k(z_i) \subset \mathcal{Z} \setminus \{z_i\}$ contains its k closest neighbors.

The mean neighbor distance is defined as

$$\rho_k(z_i) \coloneqq \frac{1}{k} \sum_{z_j \in \mathcal{N}_k(z_i)} \|z_i - z_j\|_2. \tag{8}$$

This quantity serves as a local density proxy. Smaller values correspond to denser regions of $\mathcal{Z}$, whereas larger values indicate sparser or more boundary like locations.

The statistic ρ_k is not intended as a parametric density estimator. Instead, it provides a robust geometric summary of neighborhood tightness in high dimensional spaces without introducing kernel bandwidths or smoothness assumptions.

Neighborhood structure at the dataset level is summarized by distributional statistics over $\{\rho_k(z_i)\}_{i=1}^N$. These include the mean and standard deviation, the coefficient of variation $\mathrm{CV} \coloneqq \sigma/\mu$, and upper tail quantiles such as the 90th percentile. Together, these summaries distinguish uniform redistribution within a shared region from persistent long range neighborhoods that arise when representations occupy separated submanifolds.

A.4 Quantile Partitions and Center and Periphery Subsets

Several experiments require selecting training samples from dense core regions and sparse boundary regions of the empirical manifold. Center and periphery subsets are defined using fixed quantile thresholds over $\{\rho_k(z_i)\}$. Let Q_α denote the α quantile of the empirical distribution of ρ_k on $\mathcal{Z}$. Given two fixed quantiles $0 < \alpha_\mathrm{c} < \alpha_\mathrm{p} < 1$,

$$\begin{aligned} \mathcal{D}_{\text{center}} &\coloneqq \{x_i \in \mathcal{D} : \rho_k(z_i) \le Q_{\alpha_\mathrm{c}}\}, \\ \mathcal{D}_{\text{periph}} &\coloneqq \{x_i \in \mathcal{D} : \rho_k(z_i) \ge Q_{\alpha_\mathrm{p}}\}. \end{aligned} \tag{9}$$

The thresholds $(\alpha_\mathrm{c}, \alpha_\mathrm{p})$ are treated as fixed constants across checkpoints and training settings unless stated otherwise. This construction yields paired subsets with identical formatting and supervision structure but different geometric support, allowing the effect of location in representation space to be isolated under controlled training conditions.

A.5 Checkpoints and Joint Projections

Representations are compared across a set of training checkpoints, denoted by $\mathcal{T}$. For each $t \in \mathcal{T}$, the corresponding representations $\mathcal{Z}^{(t)}$ are analyzed using quantitative statistics in $\mathbb{R}^d$ together with a shared low dimensional visualization.

Joint visualizations are constructed by fitting a single projection function $\phi : \mathbb{R}^d \to \mathbb{R}^2$ on the union $\bigcup_{t \in \mathcal{T}} \mathcal{Z}^{(t)}$. The same projection is then applied across all checkpoints. This shared fitting strategy reduces variation introduced by refitting, so differences across checkpoints reflect changes in representation geometry rather than visualization artifacts.

In later sections, sampling is carried out in a fixed reference embedding space, while geometric outcomes are evaluated in the decoder only representation space defined here.

B Additional Results on Language Retrieval Margin

Table 3 summarizes cosine margin statistics for language identification, computed using centroid based representations, across the baseline model and checkpoints E1 through E19. Reported values include the mean and standard deviation of the margin over the full test set, as well as results for individual language subsets (EN, ZH, HI, AR).

Table 3: **Centroid-based cosine margin for language retrieval across checkpoints.** Mean ± standard deviation of the margin over the full test set and per-language subsets. The test set contains 2000 samples (500 per language) and uses 4 language centroids.

Checkpoint	Overall	English	Chinese	Hindi	Arabic	n_{test}	$n_{\text{centroids}}$
Baseline	0.705±0.077	0.675±0.046	0.662±0.104	0.754±0.058	0.730±0.043	2000	4
E1	0.702±0.074	0.684±0.059	0.689±0.102	0.725±0.071	0.710±0.041	2000	4
E2	0.687±0.075	0.664±0.070	0.681±0.098	0.709±0.071	0.693±0.041	2000	4
E3	0.671±0.082	0.626±0.087	0.668±0.094	0.706±0.074	0.685±0.040	2000	4
E4	0.662±0.087	0.599±0.090	0.660±0.093	0.707±0.074	0.683±0.039	2000	4
E5	0.653±0.092	0.577±0.089	0.647±0.093	0.707±0.076	0.681±0.039	2000	4
E6	0.644±0.093	0.561±0.088	0.639±0.095	0.701±0.075	0.673±0.037	2000	4
E7	0.641±0.095	0.552±0.085	0.634±0.097	0.703±0.075	0.675±0.037	2000	4
E8	0.638±0.097	0.544±0.083	0.632±0.098	0.702±0.075	0.673±0.036	2000	4
E9	0.637±0.097	0.540±0.080	0.629±0.098	0.704±0.074	0.675±0.035	2000	4
E10	0.635±0.098	0.537±0.079	0.627±0.098	0.703±0.073	0.674±0.035	2000	4
E11	0.633±0.098	0.533±0.078	0.625±0.098	0.701±0.072	0.672±0.034	2000	4
E12	0.634±0.098	0.534±0.077	0.625±0.098	0.703±0.072	0.673±0.034	2000	4
E13	0.632±0.098	0.531±0.076	0.623±0.099	0.701±0.072	0.671±0.034	2000	4
E14	0.631±0.098	0.530±0.076	0.622±0.099	0.700±0.072	0.670±0.034	2000	4
E15	0.630±0.098	0.530±0.076	0.622±0.100	0.700±0.071	0.670±0.034	2000	4
E16	0.630±0.098	0.530±0.076	0.622±0.100	0.700±0.071	0.670±0.034	2000	4
E17	0.630±0.098	0.530±0.075	0.622±0.100	0.700±0.071	0.670±0.034	2000	4
E18	0.630±0.098	0.530±0.075	0.622±0.100	0.700±0.071	0.670±0.034	2000	4
E19	0.630±0.098	0.530±0.075	0.622±0.100	0.700±0.071	0.670±0.034	2000	4