Position: Beyond Euclidean – Foundation Models Should Embrace Non-Euclidean Geometries

Anonymous Author(s)
Affiliation
Address
email

Abstract

In the era of foundation models and Large Language Models (LLMs), Euclidean space has been the de facto geometric setting for machine learning architectures. However, recent literature has demonstrated that this choice comes with fundamental limitations. At a large scale, real-world data often exhibits inherently non-Euclidean structures, such as multi-way relationships, hierarchies, symmetries, and non-isotropic scaling, in a variety of domains, such as languages, vision, and the natural sciences. It is challenging to effectively capture these structures within the constraints of Euclidean spaces. This position paper argues that moving beyond Euclidean geometry is not merely an optional enhancement but a necessity to maintain the scaling law for the next-generation of foundation models. By adopting these geometries, foundation models could more efficiently leverage the aforementioned structures. Task-aware adaptability that dynamically reconfigures embeddings to match the geometry of downstream applications could further enhance efficiency and expressivity. Our position is supported by a series of theoretical and empirical investigations of prevalent foundation models. Finally, we outline a roadmap for integrating non-Euclidean geometries into foundation models, including strategies for building geometric foundation models via fine-tuning, training from scratch, and hybrid approaches.

1 Introduction

Foundation models, such as Large Language Models (LLMs), have emerged as a cornerstone of current AI advancements due to their ability to generalize across diverse tasks with minimal fine-tuning [15, 36, 22, 115]. Euclidean geometry has been the default framework for designing such models, largely driven by the natural compatibility of Euclidean geometry with fundamental neural network operations—such as linear transformations, convolutions, and attention mechanisms—which can be executed efficiently using standard linear algebra in Euclidean space. However, **real-world datasets often exhibit implicit non-Euclidean structures**, such as the hierarchical organization of natural language—including concept taxonomies and entailment relationships [104, 136, 82]—as well as hierarchical relationships among object classes, scenes, and their constituent categories in visual data [51, 107]. Furthermore, non-Euclidean characteristics are inherent in biological data, such as protein structures [143] and RNA-seq data [76]. Given the non-Euclidean characteristics of training data, along with the challenges faced by

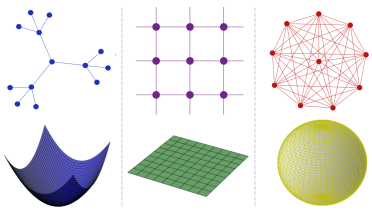


Figure 1: Manifolds with corresponding graph structures or underlying relationships, which represent different types of token relationships: hierarchical (left), uniform (middle), and cyclical (right) dependencies.

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

current foundation models—from hallucinations to computational inefficiencies—it becomes crucial to question whether Euclidean geometry should remain the default for foundation models.

Position: The development of non-Euclidean foundation models is essential for effectively representing, modeling, and analyzing complex data structures and relationships in real-world applications. Particularly, this paper advocates for the development of ***non-Euclidean foundation models*** at the scale of billions of parameters, which is on both a much larger scale and a much broader geometric scope than existing research that focuses almost entirely on low-dimensional settings within specific geometries, such as the hyperbolic space. With arguments grounded in theoretical insights and experimental evidence, we contend that by aligning foundation models—whether visual, linguistic, or scientific—with the intrinsic geometries of their training data, we can improve three critical aspects of these models: representational capabilities, adaptability to diverse geometric structures, and scalability.

Representational Capabilities. Euclidean space has limited capabilities to represent complex geometric structures with diverse local properties, resulting in significant distortion when embedding such data in low-dimensional Euclidean spaces [135]. In contrast, hyperbolic spaces, with their negative curvature, excel at representing hierarchical structures with minimal distortion in low dimensions [80, 120]. Similarly, spherical geometries, defined by positive curvature, are well-suited for modeling data with bounded structures and angular relationships [40, 41, 138].

Adaptability. Incorporating non-Euclidean geometrical operations into foundation models provides substantially enhanced adaptability to the diverse geometric structures in training datasets, particularly in the case of large-scale datasets—as is typical for these models—where heterogeneity is inherent by design. This adaptability improves the models' flexibility and transferability, as many instances of downstream tasks benefit from acknowledging the geometric structure in the data, such as knowledge graph completion [7, 145], social network analysis [156, 72], multi-label classification, drug discovery [111], user preferences recommendation [28, 155, 154], and code understanding [132].

Scalability. Adapting foundation models to non-Euclidean geometry enables expressive lower-dimensional embeddings, reducing computational costs without sacrificing performance. A critical implication lies in the scaling laws of foundation models [65], where performance in Euclidean models follows a power-law scaling of the form $L(N) \propto N^{-\alpha}$, with L and N being the loss and parameter count. This behavior reveals inherent inefficiencies in how Euclidean space handles increasing model complexity and data dimensionality. In contrast, Riemannian methods have shown promises to improve scaling by efficiently compressing information [20, 74]. For instance, hyperbolic spaces better captures long-range dependencies [136] and mixed-curvature approaches [55, 145] allow different model components to scale according to their optimal geometric properties.

Roadmap. Moreover, we propose a roadmap for integrating non-Euclidean geometries into foundation models. This includes both adapting existing Euclidean models to incorporate these principles and developing foundation models from scratch. We also highlight key challenges and outline the steps required to advance this non-Euclidean vision, from architectural design to the creation of non-Euclidean libraries, given that existing frameworks such as DeepSpeed [116] and Flash Attention [33] are tailored exclusively for Euclidean models.

2 Background and Preliminaries

In this section, we give an overview of non-Euclidean spaces, particularly focusing on Riemannian manifolds. For more details please see [84] and Appendix A.

2.1 Non-Euclidean Geometry Foundations

Riemannian Manifolds. A smooth n-dimensional manifold $\mathcal{M}$ is a topological space that is locally Euclidean. Each point x is associated with a *tangent space* $T_x\mathcal{M}$, which is an n-dimensional vector space that acts as a first-order local approximation of $\mathcal{M}$. A *Riemannian metric* $\mathfrak{g}$ on $\mathcal{M}$ is a collection $\mathfrak{g} := (\mathfrak{g}_x)_{x\in\mathcal{M}}$ of positive definite bilinear forms $\mathfrak{g}_x(\cdot,\cdot) : T_x\mathcal{M} \times T_x\mathcal{M} \to \mathbb{R}^n$, varying smoothly with x. $\mathfrak{g}_x$ induces the *(sectional) curvature* at point x, which measures how $\mathcal{M}$ deviates from flatness at x. A *Riemannian manifold* is a pairing $(\mathcal{M}, \mathfrak{g})$. For example, $\mathbb{R}^n$ with the usual Euclidean inner product is a Riemannian manifold with constant curvature 0. $\mathfrak{g}_x$ can be seen as a generalization of inner products, where the norm of $p \in T_x\mathcal{M}$ is $\|p\|_{\mathfrak{g}} = \sqrt{\mathfrak{g}_x(p,p)}$. The choice of $\mathfrak{g}$ induces a global

distance function $d(\cdot, \cdot)$ on $\mathcal{M}$. A *geodesic* between x, y is a local distance minimizing smooth curve. In particular, the shortest paths are geodesics. With certain assumption on the structure of $\mathcal{M}$, one can define the *exponential map* $\exp_x : T_x\mathcal{M} \to \mathcal{M}$ for $x \in \mathcal{M}$, and its inverse, the *logarithmic map* $\log_x : \mathcal{M} \to T_x\mathcal{M}$. Additionally, the *parallel transport* map $\mathrm{PT}_x(v, w)$, where $v, w \in T_x\mathcal{M}$, generalizes translation, transporting w starting at x in the direction of v with no acceleration.

2.2 Deep Learning in Non-Euclidean Spaces

Recent years have witnessed an increasing interest in extending deep learning techniques to Riemannian manifolds. We discuss several advances for designing neural networks and Transformers in non-Euclidean geometries, as well as optimization on manifolds, with more details in Appendix A.

Non-Euclidean Neural Networks. Several works have explored neural networks that leverage geodesic distance to perform neural network operations [21, 94, 17, 78]. Within hyperbolic learning, prior works have developed neural network layers [47, 123, 104, 26, 149], graph neural networks [91, 23], vision models [10, 141], and residual neural networks [63]. In addition, extensive works have developed equivariant neural networks that encode spherical geometry as inductive bias [31, 41, 32, 34, 42]. Neural networks for mixed curvature manifolds that encompass both hyperbolic and spherical models have also been proposed [55, 6]. Many Euclidean convex and stochastic optimization algorithms have been extended to manifold learning as well [139, 162, 11, 147, 148].

Non-Euclidean Transformers. Significant advancements have been made toward Transformers in non-Euclidean spaces in recent studies. Prior works have developed attention mechanisms and additional essential operations, such as layer normalization, to develop Transformers in hyperbolic, spherical, and mixed curvature manifolds [56, 26, 123, 159, 81, 29].

Nevertheless, there is **a lack of works for non-Euclidean foundation models**. These prior works almost all focus on low-dimensional settings, with few works that consider pre-trained models [27].

3 Foundation Models Should Embrace Non-Euclidean Geometries

Euclidean Foundation Models. Foundation models are typically trained on massive corpora to learn transferable representations that serve as a basis for downstream tasks [15]. Transformer-based language models [36, 22, 22, 113, 53], large-scale vision models such as Vision Transformer (ViT) and ResNet [39, 61], and multimodal foundation models like CLIP [112] and DALL-E [115], have achieve state-of-the-art performances in a vast amount of tasks across numerous domains.

3.1 Limitations of Euclidean Geometry for Foundation Models

The Euclidean assumption is that relationships between data points can be meaningfully characterized using distances measured in a flat space. However, theoretical and experimental works have demonstrated that **Euclidean geometry, with its isotropic nature and uniform scaling, fails to capture the complex structures of real-world data, resulting in significant distortions** [19, 95, 3, 58, 96]. As a result, high-quality, low-distortion embeddings are often only possible in *high-dimensional* Euclidean space. Specifically, embeddings of complex structured data, such as hierarchies or trees, provably incur *high rates of distortion* [18, 95]. In this section, we highlight how the flat nature of the Euclidean space results in limitations and challenges for foundational models.

Non-Applicability of the Nash Embedding Theorem. The Nash Embedding Theorem states that any Riemannian manifold $\mathcal{M}$ of dimension n admits an isometric embedding f into $\mathbb{R}^{2n+1}$ [103], seemingly to imply that non-Euclidean spaces would only reduce the embedding dimension by half. However, the isometric embedding here is defined to preserve the Riemannian metric, meaning that it is *locally distance preserving*—the length of any path is preserved. However, for the shortest path between points $x, y \in \mathcal{M}$, its image under f is not necessarily the shortest path (i.e., Euclidean straight line) between $f(x)$ and $f(y)$. Conversely, measuring the embedding distortion is concerned with whether a map is *globally distance preserving*, or when the shortest path between x and y remains the shortest path between $f(x)$ and $f(y)$, which is defined by isometric embeddings between metric spaces. Note that **an isometric embedding between Riemannian manifolds is in general not an isometric embedding between metric spaces**.

We are concerned with global distance-preserving embeddings for foundational models, as the distance between any pair of token embeddings is crucial for model training. Thus, the Nash Embedding Theorem is not applicable since global distortion could still arise from isometric embeddings between Riemannian manifolds. For this reason, by "isometry", we refer to those between metric spaces. See Appendix A.2 for more details. As the Nash Embedding Theorem is not applicable, Euclidean embeddings suffer from several limitations, which we detail below.

Dimensionality. Euclidean space requires high dimensionality to embed complex structures with low distortion. The following theorem shows the distortion-dimension tradeoff for Euclidean embeddings even in the simple case of unweighted token relationships, in the form of complete graphs.

Theorem 3.1. *(Matoušek [97]) Let X be an n-point metric space with uniform distance 1, i.e., an unweighted complete graph with n nodes. For $\epsilon > 0$, the minimal d such that X can be embedded into $\mathbb{R}^d$ with distortion $(1+\epsilon)$ is $d = \Omega\left(\frac{\log(n)}{\epsilon^2 \log(1/\epsilon)}\right)$*

For any $p < 2$, $\epsilon^2 \log(1/\epsilon)$ tends to 0 faster than ϵ^p as $\epsilon \to 0$. As a result, Theorem 3.1 implies that d grows *near-quadratically* w.r.t. inverse distortion. Furthermore, *any* unweighted graph with n nodes can be isometrically embedded into an unweighted complete graph with n nodes. Thus Theorem 3.1 implies the same dimensionality issue for embedding any unweighted graph in Euclidean space.

Distortion. Non-trivial distortion could exist regardless of the dimension of the Euclidean space in the cases of more complex structures. The following theorem implies that a wide range of spaces cannot be isometrically embedded into Euclidean space, based on *Markov convexity* (Appendix A.3).

Theorem 3.2. *[83] Let $(X, d_X), (Y, d_Y)$ be metric spaces. For every $p \in \mathbb{N}$, denote $\Pi_p(X), \Pi_p(Y)$ the Markov p-convexity constant of X and Y respectively. Let $c_Y(X) = \inf\{\mathrm{dist}(f) : f : X \to Y\}$ denote the minimum distortion of embedding X in Y. Then $c_Y(X) \geq \frac{\Pi_p(X)}{\Pi_p(Y)}$.*

When X models hierarchical token relationships, e.g., $X = B_{2^k}$ is a complete binary tree of depth 2^k, the distortion for embedding binary trees of depth in any Euclidean space is *at least* $\Omega(1) \cdot \sqrt{\log k}$. When X represents circular or periodic dependencies in tokens, e.g., X is a ball of radius r in a vertex-transitive graph, the minimal distortion of embedding X into $\mathbb{R}^n$ for any n is $\Omega(\sqrt{\log r})$ [83].

Moreover, non-trivial distortion exists when embedding other forms of topological space as well, including the sphere $S^k \subseteq \mathbb{R}^{k+1}$, as shown in the following theorem.

Theorem 3.3. *[117] Let (X, d_X) be a metric space with $X = \{a, b, c, d\}$ and $d_X(a,b) = d_X(a,c) = d_X(a,d) = 2L$ and $d_X(b,d) = d_X(c,d) = L$ for $L \in \mathbb{R}^+$. Then X admits no isometric embedding into $\mathbb{R}^n$ for any n.*

As these points can be isometrically embedded into S^k, Theorem 3.3 shows that S^k cannot be isometrically embedded into $\mathbb{R}^n$ for any $n \in \mathbb{N}$, resulting in distortion when encoding rotational equivariance. In contrast, non-Euclidean geometry can provide a more natural representation of complex topological structures, **reducing distortion and dimensionality** of the embedding space. For instance, [120] showed that every finite tree admits an embedding into the hyperbolic plane $\mathbb{H}^2$ with $1+\epsilon$ multiplicative distortion for any $\epsilon > 0$, leading to $O(1)$ distortion with low dimensionality.

Take-away. The implications of the previous theoretical discussion are numerous: (1) **Limited scalability.** Theorem 3.1 highlights the distortion-dimension trade-off for Euclidean foundation models when embedding complex structures, which is reflected in the computational resources required in these models. Non-Euclidean geometry produces higher quality embeddings in significantly lower dimensions, offering enhanced model scalability; (2) **Performance bottleneck.** Theorem 3.2 and 3.3 demonstrate that even in the case of an abundance of compute resources, the linear assumption in Euclidean foundation models could still incur significant distortion regardless of the embedding dimension for a wide range of topological structures, resulting in a performance upper bound.

3.2 Non-Euclidean Geometry in Foundation Models

In this section, we empirically assess embedding distortions for different geometries to validate our claims in Section 3.1 and demonstrate that non-Euclidean geometry is more suitable. We then analyze token embeddings in foundation models, showing that **structures that align with non-Euclidean geometry are prevalent, highlighting the need for alternative geometric frameworks**.

Table 1: δ-Hyperbolicity of the token embedding in various LLMs across several datasets. The bottom 2 rows show the δ-hyperbolicity values of several metric spaces for reference.

Model	arXiv	C4	Common Crawl	GitHub	StackExchange	Wikipedia
RoBERTa-Base [92]	0.15 ± 0.06	0.18 ± 0.04	0.17 ± 0.04	0.12 ± 0.04	0.17 ± 0.07	0.07 ± 0.05
LLaMA3.1-8B [53]	0.15 ± 0.05	0.16 ± 0.07	0.15 ± 0.06	0.12 ± 0.05	0.18 ± 0.06	0.10 ± 0.04
GPT-NeoX-20B [14]	0.14 ± 0.03	0.17 ± 0.06	0.15 ± 0.05	0.11 ± 0.04	0.14 ± 0.04	0.09 ± 0.03
Gemma2-9B [134]	0.17 ± 0.06	0.19 ± 0.04	0.20 ± 0.05	0.15 ± 0.05	0.18 ± 0.04	0.15 ± 0.03
Metric Space	Sphere Space	Dense Graph	PubMed Graph	Poincaré Space	Tree Graph	-
Reference δ values	0.99 ± 0.01	0.63 ± 0.01	0.40 ± 0.04	0.14 ± 0.01	0.0	-

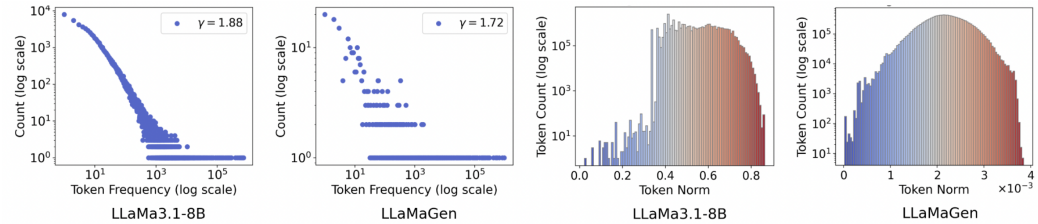


Figure 2: Token frequency v.s. token count (left 2) and token norm vs token count (right 2) for LLaMa3.1-8B and LLaMaGen. The datasets are chosen to be within the training corpus. The token-frequency figures show the scale-free properties of the token inputs. The token norms figures reflect this property for learned token embeddings to some extent, with token count decreasing exponentially for high normed tokens at the right tail. However, the Euclidean embeddings stil do not fully capture this property and deviate from it at the left tail. More statistics are shown in Appendix B.

Figure 3: Average (point-wise) distortion on canonical graphs with 96 nodes, comparing four spaces with total dimension 6. The least distortion is achieved by the space with the most suitable geometry.

Geometry	Tree $\lvert E\rvert = 95, \lvert V\rvert = 96$	Cycle $\lvert E\rvert = 96, \lvert V\rvert = 96$	Ring of Trees $\lvert E\rvert = 96, \lvert V\rvert = 96$
$\mathbb{R}^6$	0.1036	0.1042	0.1060
$\mathbb{H}^{-1,6}$	**0.0454**	0.2356	0.0736
$\mathbb{S}^{1,6}$	0.1440	**0.0011**	0.1365
$\mathbb{H}^{-1,3} \times \mathbb{S}^{1,3}$	0.0624	0.1337	**0.0686**

Empirical Validation. We empirically validate our claim that Euclidean space fails to capture complex structures faithfully and that non-Euclidean spaces are better suited for producing high-quality embeddings. Table 3 compares the average (point-wise) distortion of four geometric spaces ($\mathbb{R}^6$, $\mathbb{H}^{-1,6}$, $\mathbb{S}^{1,6}$, and $\mathbb{H}^{-1,3} \times \mathbb{S}^{1,3}$) in representing three canonical graphs (Tree, Cycle, and Ring of Trees) with 96 nodes, each corresponding to a different type of intrinsic token relationships (hierarchical, cyclical, and both). The most suitable geometry varies by graph type—Lorentzian space ($\mathbb{H}^{-1,6}$) for trees, spherical space ($\mathbb{S}^{1,6}$) for cycles, and mixed geometry ($\mathbb{H}^{-1,3} \times \mathbb{S}^{1,3}$) for rings of trees—emphasizing the importance of selecting an appropriate geometry to minimize distortion.

We also compute the distortion value against varying dimensionality. An example is shown in Figure 4 for the case of a tree with 96 nodes, plotted on log-scale for visibility. The hybrid manifold is a product of hyperbolic and spherical spaces, each with half the dimension. The 4-dimensional hyperbolic space achieves a significantly smaller distortion than Euclidean embeddings with 50 dimensions. **This reflects Takeaway 1 in Section 3.1**, where non-Euclidean geometry achieves superior performance with significantly fewer dimensions. Additionally, distortion continues to decrease for hyperbolic and hybrid spaces but plateaus for Euclidean space, **reflecting Takeaway 2**, where Euclidean space has theoretical upper bounds for embedding trees but non-Euclidean geometry has the potential to continue the performance scaling law at high dimensionality. See Appendix B for additional plots of other graph types.

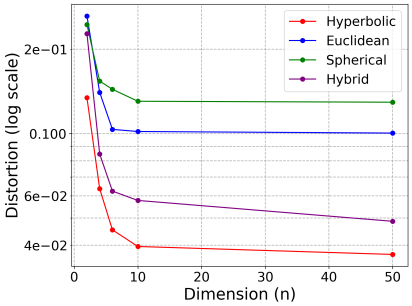


Figure 4: Distortion for embedding a Tree with 96 nodes for varying dimensionality (log scale). Non-Euclidean geometry achieves smaller distortion with significantly fewer dimensions and has better scaling.

Hierarchies in Token Embeddings. Based on the above results validating Euclidean embeddings' limitations, we further show that these structures which Euclidean spaces struggles to embed is prevalent in foundation models. To explore the intrinsic structure within the representations of foundation models, we utilize **δ-hyperbolicity** [54], which quantifies the extent to which a metric

space deviates globally from a tree metric (see Appendix C). Each token is treated as a point in a discrete metric space X, and a graph is constructed based on similarity scores between each pair. We evaluate the hyperbolicity of token embeddings in LLMs, where lower values suggest a tree-like structure. As shown in Table 1, the consistently low δ-hyperbolicity values suggest *hierarchical structures* within each prompt across diverse datasets.

We also analyze the global token embedding distribution in LLMs and pre-trained vision models using datasets included in the models' training corpus [137, 53, 130]. Figure 2 plots token input frequency distributions and their occurrences in the dataset on a log scale, revealing a *scale-free* structure among the token embeddings. This scale-free organization suggests an underlying hierarchical structure [9], where a small number of high-frequency tokens act as hubs within the semantic network. The figure also shows token norm distributions for learned embeddings, where the count for high-norm embeddings decreases exponentially at the right tail, reinforcing the scale-free property. The non-Euclidean structures in token distribution are exhibited to some extent even in Euclidean models are most likely attributed to the models being optimized during training to maximize representational quality. However, *the scale-free properties are still not yet fully captured by the Euclidean foundational model, where the count of embeddings with small norms still increases.* See Appendix B for more statistics.

Additional Structures. In addition to hierarchical structure, data may exhibit other structural characteristics, such as cycles and loops. Many real-world tasks, such as 3D shape analysis [41, 42], medical imaging [12, 152], and physics-informed machine learning [89, 90, 2, 31], can benefit from encoding data geometry as inductive bias. Euclidean operations, such as convolutional layers, encode only translation invariance [43], resulting in performance limitations for these tasks.

3.3 The Necessity of Non-Euclidean Geometry for Foundation Models

Here we further explore how non-Euclidean geometry could improve foundation model performance.

(1) Addressing the limitations in capturing intrinsic token structures. Recent research shows that the attention mechanism plays a pivotal role in the expressive capacity of LLMs [1, 124, 142, 8].

Lemma 3.4 (Balestriero et al. [8])**.** *Let $X \in \mathbb{R}^{T\times D_{(\ell)}}$ be the input to the ℓ-th layer of an LLM, where T is sequence length and $D_{(\ell)}$ is feature dimension. Attention head h's output at position i is in the convex hull of the first i rows of $XV_{h,(\ell)}$: $\mathit{Head}_{h,(\ell)}(X)_i \in \mathit{Hull}\left\{(V_{h,(\ell)})^\top x_j \mid j=1,\dots,i\right\}$. with bounded effective dimension: $\mathit{dim}_{\mathit{eff}} \le \#\left\{\mathit{Attn}_{h,(\ell)}(X)_{i,j}>0 \mid j\in\{1,\dots,i\}\right\}$. Here, $\mathit{Attn}_{h,(\ell)}(X)$ is the attention matrix for head h at layer ℓ: $\mathit{Attn}_{h,(\ell)}(X) = \mathit{softmax}_{\mathit{causal}}(XQ_{h,(\ell)}K_{h,(\ell)}^\top X^\top)$.*

This lemma *highlights that next-token prediction in LLMs is strongly influenced by relationships encoded in previous tokens.* As shown in Table 1, tokens exhibits non-Euclidean characteristics. Consequently, the standard Euclidean attention mechanism does not faithfully capture hierarchical syntax, periodic dependencies, and other complex token relationships, as demonstrated in Section 3.1. Utilizing non-Euclidean attention mechanisms instead could *better capture previous token relationships by aligning with the intrinsic data structure, thus enhancing next-token prediction.* For example, hyperbolic geometry compresses distances exponentially, ensuring that distant but structurally related tokens (e.g., a root concept and its distant co-occurrences in a prompt) remain meaningfully close, enabling attention mechanisms to *efficiently capture long-range dependencies and hierarchies.*

(2) Alleviating distortion-dimension trade-offs. Recent studies examined how Euclidean-based LLMs encode hierarchies geometrically [108, 109], where a mapping function λ maps input text x to a vector $\lambda(x) \in \mathbb{R}^d$, and an un-embedding layer assigns $\gamma(y) \in \mathbb{R}^d$ to each token y. The token probability distribution is given by $P(y \mid x) = \frac{\exp(\lambda(x)^\top \gamma(y))}{\sum_{y'\in \text{Vocab}} \exp(\lambda(x)^\top \gamma(y'))}$. To unify the different spaces, the embedding and unembedding spaces can be reformulated using transformations $g(y) = A(\gamma(y) - \bar{\gamma}_0), \quad \ell(x) = A^{-\top}\lambda(x)$, where the Euclidean inner product serves as the causal inner product. This framework shows that Euclidean LLMs encode hierarchical concepts *orthogonally*, where parent (e.g., animal) and child (e.g., bird, mammal) vectors are perpendicular. However, this orthogonal representation demands *high dimensionality*, which scales significantly as hierarchical relationships grow more complex [58]. *Non-Euclidean spaces offer a more efficient alternative, preserving hierarchical relationships while significantly reducing dimensionality [104, 105].*

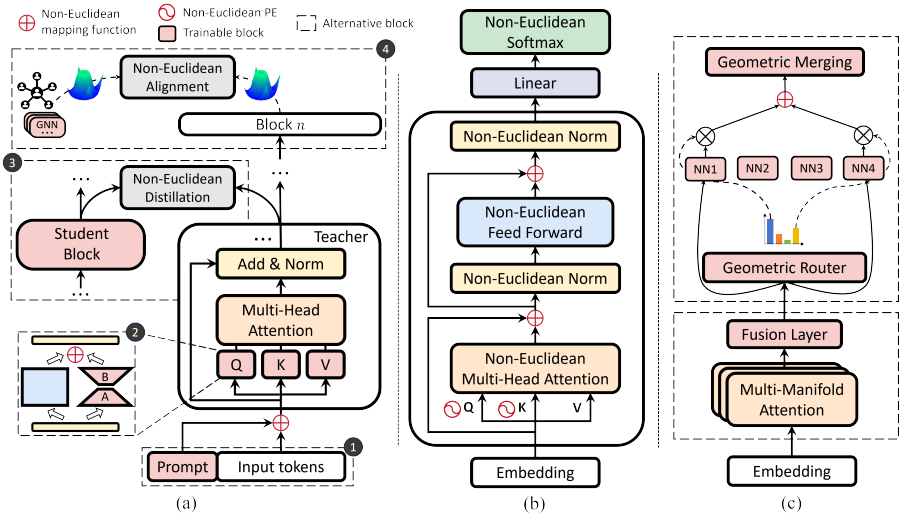


Figure 5: Roadmap for integrating non-Euclidean geometries into foundation models, includes (a) fine-tuning existing Euclidean foundation models, (b) pretraining from scratch, and (c) hybrid architectures. Four strategies are shown in (a), labeled with circled numbers 1-4, respectively: geometric prompt tuning, geometric low-rank adaptation, geometric knowledge distillation, and geometric transfer learning. All learnable components are highlighted in red in (a) and (c).

(3) Improved multi-modal heterogeneity modeling. Data from different modalities vary significantly due to contextual factors, use cases, cultural differences, and different interpretations of the same information. This complexity intensifies in multi-modal data, where each modality has distinct complex structures [87, 150, 88, 52, 68]. For instance, latent modality gap and distinct modality structures exist in the latent space due to initialization and the contrastive learning process, impacting downstream tasks [88]. Different modalities also exist on separate manifolds [144], making a unified Euclidean foundation model *highly redundant in parameters and requiring varying degrees of pruning for different modalities.* Thus, Euclidean space struggles to capture multi-modal cross-domain relationships, as its flat structure lacks the flexibility needed for multi-faceted interactions in the data.

Non-Euclidean spaces exhibit much more geometric flexibility to enable multiple manifolds that encode different data distributions [151, 49, 48]. For instance, hyperbolic geometry excels in vision-language foundation models by effectively capturing hierarchical relationships [35, 107, 114], improving performance in tasks such as image-video-skeleton [86] and video-audio applications [66] while enhancing representation interpretability—higher-level hierarchical concepts lie closer to the origin with more specific concepts residing in more peripheral regions, enabling geodesic reasoning when navigating through concept hierarchies.

4 Towards Non-Euclidean Foundation Models

We propose a roadmap that explores three progressive approaches to incorporate non-Euclidean geometry in foundation models: fine-tuning existing Euclidean models, building non-Euclidean models from scratch, and developing a hybrid framework combining both for optimal performance. We also discuss key implementation challenges of the roadmap.

4.1 Fine-tuning Existing Euclidean Foundation Models

Off-the-shelf pre-trained Euclidean foundation models are strong starting points as they already encode rich information. An efficient strategy is to adapt them to non-Euclidean spaces, thereby retaining their original capabilities and enabling generalization to data with non-Euclidean structures. We propose four strategies, shown in Figure 5(a): (1) Geometric prompt tuning; (2) Geometric low-rank adaptation; (3) Geometric knowledge distillation; and (4) Geometric transfer learning.

(1) Geometric Prompt Tuning. Prompt tuning offers a parameter-efficient alternative to full fine-tuning by introducing trainable, task-specific prompt tokens to the input, mitigating catastrophic

forgetting while requiring fewer trainable parameters. [75, 59]. Geometric prompts can be optimized through non-Euclidean spaces to better align with the data geometry to and adapt to downstream tasks. For instance, trainable prompt and token embeddings could be introduced to better capture the topological relationships between prompts and text inputs.

(2) Geometric Low-Rank Adaptation. Low-Rank Adaptation (LoRA) offers an efficient way to adjust the model parameter space for downstream tasks [67]. To equip the pre-trained model with non-Euclidean geometry through geometric low-rank adaptation, low-rank matrix multiplications could be performed directly on the manifold after projecting the input into non-Euclidean spaces, which better models the underlying geometric structure of the data [157].

(3) Geometric Knowledge Distillation. Distilling knowledge into non-Euclidean spaces refers to transferring knowledge from a large, complex teacher model to a smaller, more efficient student model by utilizing manifold properties to teach the student to better inherit the teacher model's geometric structure. An example is minimizing the gap between each layer's output of both models, especially in high-dimensional spaces [153, 60] and resource-limited applications.

(4) Geometric Transfer Learning. Geometric transfer learning aims to help foundation models learn across domains with aligned geometries, ensuring a much more effective and consistent knowledge transfer. Geometry alignment objectives can be designed to supervise the transfer of geometric knowledge, such as hyperbolic contrastive learning for recommendation [160, 93], preserving the intrinsic structure of the target domain while retaining geometry-agnostic prior knowledge.

4.2 Pretraining from Scratch

Pretraining non-Euclidean foundation models requires addressing unique challenges. We outline key components for adapting models to complex curvature-aware structures; see also Figure 5(b). A detailed mathematical formulation is presented in Table 3 in the Appendix.

Curvature Estimation. A manifold's curvature determines its intrinsic geometric properties, such as distance metrics and learning dynamics. Curvature estimation methods vary based on data types. For *graph data* (e.g., networks, proteins), curvature can be derived from topological properties, such as Ollivier-Ricci curvature or Gromov hyperbolicity [106, 69, 157, 50]. For *non-graph data* (e.g., texts, images), curvature can be estimated from learned embeddings [73, 4] or techniques like Isomap [135] and UMAP [99]. One could also design learnable curvature within training pipelines using second-order statistics [49], reinforcement learning [46], and self-supervised learning [128, 127].

Non-Euclidean Attention Mechanism. In non-Euclidean spaces, attention scores can be defined based on negative manifold distance $-d_{\mathcal{M}}(x, y)$ between queries and keys instead of dot products [56, 123, 26, 29], with closer node pairs receiving higher attention weights. To aggregate attention, unified manifold centroids or tangent space operations can be used [55, 29]. Linear attention mechanisms [159] can be employed to improve computational efficiency by approximating traditional attention through unified tangent space operations.

Other Important Modules. Traditional Euclidean positional encodings [142, 126] do not preserve the manifold structure in non-Euclidean spaces. Several approaches for non-Euclidean positional encoding [26, 159, 44] were proposed to represent token positions while maintaining geometric integrity. *Residual connections* should be formulated using isometric operations [63, 141, 10] to preserve geometric information across layers. Layer and batch normalization must also be adapted to account for curvature [159, 10, 141]. Loss function must also satisfy geometric constraints, such as computing the probability distribution over tokens based on the manifold distance instead.

4.3 Hybrid Architectures

Hybrid architectures take a step further by merging both Euclidean and non-Euclidean foundation model architectures to provide a more universal inductive bias. We illustrate two promising strategies, also depicted in Figure 5(c).

Dynamic Geometry Adaptation. An intuitive way for hybrid modeling is to design an efficient and geometry-aware mechanism that shifts dynamically between manifolds. Unified product manifold frameworks [129] could enable layers to integrate diverse learnable curvature values that adapt to fine-grained geometric structures. Mixture of Experts (MoE) [165] provides a natural framework for

hybrid paradigms to use a geometry-aware sparse routing network by selecting the most appropriate geometry considering input structure [57], addressing issues of distortion and heterogeneity.

Multi-Manifold Attention. Multi-manifold attention could lead to more versatile underlying dependencies [79, 70], where the input is embedded into a collection of manifolds (including Euclidean) to represent differences in geometric structure across the dataset. These geometric attention maps are then fused to produce a highly discriminative map for improved attention guidance.

4.4 Roadmap Implementation Challenges

One concern for non-Euclidean foundation models is that non-Euclidean operations often more computationally intensive. In particular, tangent-space-enabled methods [47, 23, 141], although intuitive by taking advantage of the Euclidean structure of the tangent spaces, incur significant computation overhead due to multiple mappings to and from the tangent bundle. In comparison, methods that operate directly on the manifold [26, 159, 63], while still computationally more expensive than Euclidean methods, typically have similar computational complexity as their Euclidean counterparts. This type of operation could be promising for managing the computational efficiency of non-Euclidean foundation models. Additionally, non-Euclidean models require fewer dimensions to embed complex structures, as seen in our discussion in Section 3.1 and 3.2. This enables the potential for non-Euclidean models to match the performance of Euclidean models with fewer parameters to offset the computational overhead while offering additional benefits, such as the potential to continue the scaling law relationship between parameters and model performance.

Additional challenges include gaps in current research for concrete analysis between embedding quality and model performance. Some previous works have shown that using manifolds that more accurately estimate the structure of the data could enhance performance in graph tasks and word embeddings [55]. However, to the best of our knowledge, there currently lacks conclusive work connecting distortion to downstream performance, which is challenging as it could require prior knowledge of the ground-truth data geometry, compute resources to train multiple foundation models, and isolating the effects of distortion. Future works in this aspect would provide valuable insights to support better development of non-Euclidean methods for foundation models.

5 Alternative Views

While non-Euclidean geometries have clear theoretical benefits, non-Euclidean operations add complexity, which might reduce some of the anticipated efficiency benefits. In addition to points mentioned in Section 4.4, it is essential to develop libraries, such as [62] that optimize these computations, with efficient implementations of tensor operations that encode the underlying geometry.

Another view is that as hardware capacities increase, simply scaling Euclidean models to higher dimensions might reduce distortion. However, as pointed out by Theorem 3.2 and 3.3, there is an upper bound in how much distortion reduction is possible. Previous works have also empirically shown that non-Euclidean models outperform Euclidean models even with scaled parameter counts, such as for equivariant and non-equivariant models [20]. Additionally, in many domains, such as molecular structures or rare languages, data scarcity results in brute-force scaling being ineffective. Non-Euclidean geometries, on the other hand, can capture important relationships even in lower-dimensional settings [120], making them efficient in data requirements, offering better performance scalability w.r.t. model size, and are more reliable for domains with limited high-quality data.

6 Conclusion

Foundation models benefit from embracing non-Euclidean geometry to resolve their inherent mismatch with the non-Euclidean nature of real-world data. Non-Euclidean geometries reduce distortion for embedding complex structures and relationships while enabling efficient representations, which is critical for trillion-parameter scaling. Aligning architectures with data geometry could mitigate hallucinations, boost efficiency, and unlock heterogeneous scaling. We encourage the community to consider three directions: unified curvature-adaptive foundation models, geometry-aware benchmarks, and studying manifold-emergent capability links. Embracing this paradigm will catalyze AI systems that better reflect the rich geometries of human knowledge and physical reality.

References

[1] Armen Aghajanyan, Sonal Gupta, and Luke Zettlemoyer. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. In *ACL-IJCNLP*, pages 7319–7328, 2021.

[2] Brandon Anderson, Truong-Son Hy, and Risi Kondor. Cormorant: Covariant molecular neural networks. In *NeurIPS*, 2019.

[3] Juan Arias-de Reyna and Luis Rodríguez-Piazza. Finite metric spaces needing high dimension for lipschitz embeddings in banach spaces. *Israel Journal of Mathematics*, 79:103–111, 1992.

[4] Mina Ghadimi Atigh, Julian Schoep, Erman Acar, Nanne Van Noord, and Pascal Mettes. Hyperbolic image segmentation. In *ICCV*, pages 4453–4462, 2022.

[5] Miroslav Bacák. Convex analysis and optimization in Hadamard spaces. In *Convex Analysis and Optimization in Hadamard Spaces*. de Gruyter, 2014.

[6] Gregor Bachmann, Gary Bécigneul, and Octavian Ganea. Constant curvature graph convolutional networks. In *ICML*, pages 486–496. PMLR, 2020.

[7] Yushi Bai, Zhitao Ying, Hongyu Ren, and Jure Leskovec. Modeling heterogeneous hierarchies with relation-specific hyperbolic cones. *Advances in Neural Information Processing Systems*, 34:12316–12327, 2021.

[8] Randall Balestriero, Romain Cosentino, and Sarath Shekkizhar. Characterizing large language model geometry helps solve toxicity detection and generation. In *ICML*.

[9] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999.

[10] Ahmad Bdeir, Kristian Schwethelm, and Niels Landwehr. Fully hyperbolic convolutional neural networks for computer vision. In *ICLR*, 2024.

[11] Gary Bécigneul and Octavian-Eugen Ganea. Riemannian adaptive optimization methods. *arXiv:1810.00760*, 2018.

[12] Erik J Bekkers, Maxime W Lafarge, Mitko Veta, Koen AJ Eppenhof, Josien PW Pluim, and Remco Duits. Roto-translation covariant convolutional networks for medical image analysis. *MICCAI*, 2018.

[13] Nithya Bhasker, Hattie Chung, Louis Boucherie, Vladislav Kim, Stefanie Speidel, and Melanie Weber. Contrastive poincaré maps for single-cell data analysis. In *ICLR 2024 Workshop on Machine Learning for Genomics Explorations*, 2024.

[14] Sid Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, Michael Pieler, USVSN Sai Prashanth, Shivanshu Purohit, Laria Reynolds, Jonathan Tow, Ben Wang, and Samuel Weinbach. GPT-NeoX-20B: An open-source autoregressive language model. In *Proceedings of the ACL Workshop on Challenges & Perspectives in Creating Large Language Models*, 2022.

[15] R. Bommasani, D. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, et al. On the opportunities and risks of foundation models. *arXiv*, 2021.

[16] Silvere Bonnabel. Stochastic gradient descent on Riemannian manifolds. *IEEE Transactions on Automatic Control*, 58(9):2217–2229, 2013.

[17] Davide Boscaini, Jonathan Masci, Emanuele Rodolà, and Michael M. Bronstein. Learning shape correspondence with anisotropic convolutional neural networks. In *NeurIPS*, 2016.

[18] Jean Bourgain. On lipschitz embedding of finite metric spaces in hilbert space. *Israel Journal of Mathematics*, 52:46–52, 1985.

[19] Jean Bourgain. The metrical interpretation of superreflexivity in banach spaces. *Israel Journal of Mathematics*, 56:222–230, 1986.

[20] Johann Brehmer, Sönke Behrends, Pim de Haan, and Taco Cohen. Does equivariance matter at scale? *arXiv:2410.23179*, 2024.

[21] 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.

[22] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *NeurIPS*, volume 33, pages 1877–1901, 2020.

[23] Ines Chami, Zhitao Ying, Christopher Ré, and Jure Leskovec. Hyperbolic graph convolutional neural networks. In *NeurIPS*, pages 4868–4879, 2019.

[24] Boli Chen, Yao Fu, Guangwei Xu, Pengjun Xie, Chuanqi Tan, Mosha Chen, and Liping Jing. Probing bert in hyperbolic spaces. *arXiv:2104.03869*, 2021.

[25] Jinghong Chen, Chong Zhao, Qicong Wang, and Hongying Meng. Hmanet: Hyperbolic manifold aware network for skeleton-based action recognition. *IEEE Transactions on Cognitive and Developmental Systems*, 15(2):602–614, 2022.

[26] Weize Chen, Xu Han, Yankai Lin, Hexu Zhao, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. Fully hyperbolic neural networks. *arXiv:2105.14686*, 2021.

[27] Weize Chen, Xu Han, Yankai Lin, Kaichen He, Ruobing Xie, Jie Zhou, and Zhiyuan Liu. Hyperbolic pre-trained language model. *IEEE TASLP*, 32, 2024.

[28] Yankai Chen, Menglin Yang, Yingxue Zhang, Mengchen Zhao, Ziqiao Meng, Jianye Hao, and Irwin King. Modeling scale-free graphs with hyperbolic geometry for knowledge-aware recommendation. In *Proceedings of the fifteenth ACM international conference on web search and data mining*, pages 94–102, 2022.

[29] Sungjun Cho, Seunghyuk Cho, Sungwoo Park, Hankook Lee, Honglak Lee, and Moontae Lee. Curve your attention: Mixed-curvature transformers for graph representation learning. *arXiv:2309.04082*, 2023.

[30] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv:2110.14168*, 2021.

[31] Taco S Cohen, Mario Geiger, Jonas Köhler, and Max Welling. Spherical cnns. *arXiv:1801.10130*, 2018.

[32] Benjamin Coors, Alexandru Paul Condurache, and Andreas Geiger. Spherenet: Learning spherical representations for detection and classification in omnidirectional images. In *ECCV*, 2018.

[33] Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. FlashAttention: Fast and memory-efficient exact attention with IO-awareness. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[34] Michaël Defferrard, Martino Milani, Frédérick Gusset, and Nathanaël Perraudin. DeepSphere: a graph-based spherical CNN. In *ICLR*, 2020.

[35] Karan Desai, Maximilian Nickel, Tanmay Rajpurohit, Justin Johnson, and Shanmukha Ramakrishna Vedantam. Hyperbolic image-text representations. In *ICML*, pages 7694–7731. PMLR, 2023.

[36] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv:1810.04805*, 2018.

[37] Bhuwan Dhingra, Christopher J Shallue, Mohammad Norouzi, Andrew M Dai, and George E Dahl. Embedding text in hyperbolic spaces. *arXiv:1806.04313*, 2018.

[38] Jiarui Ding and Aviv Regev. Deep generative model embedding of single-cell rna-seq profiles on hyperspheres and hyperbolic spaces. *Nature communications*, 12(1):2554, 2021.

[39] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv:2010.11929*, 2020.

[40] Aiden Durrant and Georgios Leontidis. Hyperspherically regularized networks for self-supervision. *Image and Vision Computing*, 124:104494, 2022.

[41] Carlos Esteves, Christine Allen-Blanchette, Ameesh Makadia, and Kostas Daniilidis. Learning so(3) equivariant representations with spherical cnns. *CoRR*, 2017.

[42] Carlos Esteves, Yinshuang Xu, Christine Allen-Blanchette, and Kostas Daniilidis. Equivariant multi-view networks. In *The IEEE International Conference on Computer Vision*, October 2019.

[43] Carlos Esteves, Ameesh Makadia, and Kostas Daniilidis. Spin-weighted spherical cnns. *CoRR*, 2020.

[44] Jacob Fein-Ashley, Ethan Feng, and Minh Pham. Hvt: A comprehensive vision framework for learning in non-euclidean space. *arxiv*, 2024.

[45] Oluwaseyi Feyisetan, Tom Diethe, and Thomas Drake. Leveraging hierarchical representations for preserving privacy and utility in text. In *ICDM*, pages 210–219. IEEE, 2019.

[46] Xingcheng Fu, Jianxin Li, Jia Wu, Qingyun Sun, Cheng Ji, Senzhang Wang, Jiajun Tan, Hao Peng, and S Yu Philip. Ace-hgnn: Adaptive curvature exploration hyperbolic graph neural network. In *ICDM*, pages 111–120. IEEE, 2021.

[47] Octavian Ganea, Gary Bécigneul, and Thomas Hofmann. Hyperbolic neural networks. In *NeurIPS*, pages 5345–5355, 2018.

[48] Zhi Gao, Yuwei Wu, Yunde Jia, and Mehrtash Harandi. Curvature generation in curved spaces for few-shot learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8691–8700, 2021.

[49] Zhi Gao, Yuwei Wu, Mehrtash Harandi, and Yunde Jia. Curvature-adaptive meta-learning for fast adaptation to manifold data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):1545–1562, 2022.

[50] Nicolas Garcia Trillos and Melanie Weber. Continuum limits of ollivier's ricci curvature on data clouds: pointwise consistency and global lower bounds. *arXiv:2307.02378*, 2023.

[51] Songwei Ge, Shlok Mishra, Simon Kornblith, Chun-Liang Li, and David Jacobs. Hyperbolic contrastive learning for visual representations beyond objects. In *ICCV*, pages 6840–6849, 2023.

[52] Shashank Goel, Hritik Bansal, Sumit Bhatia, Ryan Rossi, Vishwa Vinay, and Aditya Grover. Cyclip: Cyclic contrastive language-image pretraining. *Advances in Neural Information Processing Systems*, 35:6704–6719, 2022.

[53] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv:2407.21783*, 2024.

[54] Mikhael Gromov. *Hyperbolic groups*. Springer, 1987.

[55] Albert Gu, Frederic Sala, Beliz Gunel, and Christopher Ré. Learning mixed-curvature representations in product spaces. *ICLR*, 2019.

[56] Caglar Gulcehre, Misha Denil, Mateusz Malinowski, Ali Razavi, Razvan Pascanu, Karl Moritz Hermann, Peter Battaglia, Victor Bapst, David Raposo, Adam Santoro, et al. Hyperbolic attention networks. In *ICLR*, 2019.

[57] Zihao Guo, Qingyun Sun, Haonan Yuan, Xingcheng Fu, Min Zhou, Yisen Gao, and Jianxin Li. Graphmore: Mitigating topological heterogeneity via mixture of riemannian experts, 2024.

[58] Anupam Gupta. Embedding tree metrics into low dimensional euclidean spaces. In *Proceedings of the thirty-first annual ACM symposium on Theory of computing*, pages 694–700, 1999.

[59] Zeyu Han, Chao Gao, Jinyang Liu, Jeff Zhang, and Sai Qian Zhang. Parameter-efficient fine-tuning for large models: A comprehensive survey. *arXiv:2403.14608*, 2024.

[60] 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. In *NeurIPS*, volume 35, pages 9164–9175, 2022.

[61] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[62] Neil He, Menglin Yang, and Rex Ying. Hypercore: The core framework for building hyperbolic foundation models with comprehensive modules. *arXiv preprint arXiv:2504.08912*, 2025.

[63] Neil He, Menglin Yang, and Rex Ying. Lorentzian residual neural networks. In *KDD*, 2025.

[64] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *NeurIPS*, 2021.

[65] 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, Oriol Vinyals, Jack W. Rae, and Laurent Sifre. Training compute-optimal large language models. In *NeurIPS*, Red Hook, NY, USA, 2022. ISBN 9781713871088.

[66] Jie Hong, Zeeshan Hayder, Junlin Han, Pengfei Fang, Mehrtash Harandi, and Lars Petersson. Hyperbolic audio-visual zero-shot learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7873–7883, 2023.

[67] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv:2106.09685*, 2021.

[68] Qian Jiang, Changyou Chen, Han Zhao, Liqun Chen, Qing Ping, Son Dinh Tran, Yi Xu, Belinda Zeng, and Trishul Chilimbi. Understanding and constructing latent modality structures in multi-modal representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7661–7671, 2023.

[69] Jürgen Jost and Shiping Liu. Ollivier's ricci curvature, local clustering and curvature-dimension inequalities on graphs. *Discrete & Computational Geometry*, 51:300–322, 2014.

[70] Georgios Kalitsios, Dimitrios Konstantinidis, Petros Daras, and Kosmas Dimitropoulos. Dynamic grouping with multi-manifold attention for multi-view 3d object reconstruction. *IEEE Access*, 2024.

[71] Tejaswi Kasarla, Gertjan Burghouts, Max Van Spengler, Elise Van Der Pol, Rita Cucchiara, and Pascal Mettes. Maximum class separation as inductive bias in one matrix. In *NeurIPS*, volume 35, pages 19553–19566, 2022.

[72] W Sean Kennedy, Onuttom Narayan, and Iraj Saniee. On the hyperbolicity of large-scale networks. *arXiv:1307.0031*, 2013.

[73] Valentin Khrulkov, Leyla Mirvakhabova, Evgeniya Ustinova, Ivan Oseledets, and Victor Lempitsky. Hyperbolic image embeddings. In *ICCV*, pages 6418–6428, 2020.

[74] Bobak Kiani, Jason Wang, and Melanie Weber. Hardness of learning neural networks under the manifold hypothesis. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[75] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.

[76] Anna Klimovskaia, David Lopez-Paz, Léon Bottou, and Maximilian Nickel. Poincaré maps for analyzing complex hierarchies in single-cell data. *Nature communications*, 11(1):2966, 2020.

[77] Rik Koncel-Kedziorski, Subhro Roy, Aida Amini, Nate Kushman, and Hannaneh Hajishirzi. MAWPS: A math word problem repository. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1152–1157. Association for Computational Linguistics, 2016.

[78] Lecheng Kong, Yixin Chen, and Muhan Zhang. Geodesic graph neural network for efficient graph representation learning. In *NeurIPS*, 2022.

[79] Dimitrios Konstantinidis, Ilias Papastratis, Kosmas Dimitropoulos, and Petros Daras. Multi-manifold attention for vision transformers. *IEEE Access*, 2023.

[80] Dmitri Krioukov, Fragkiskos Papadopoulos, Maksim Kitsak, Amin Vahdat, and Marián Boguná. Hyperbolic geometry of complex networks. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, 82(3):036106, 2010.

[81] Xin Lai, Yukang Chen, Fanbin Lu, Jianhui Liu, and Jiaya Jia. Spherical transformer for lidar-based 3d recognition. In *CVPR*, 2023.

[82] Matt Le, Stephen Roller, Laetitia Papaxanthos, Douwe Kiela, and Maximilian Nickel. Inferring concept hierarchies from text corpora via hyperbolic embeddings. *arXiv:1902.00913*, 2019.

[83] James R. Lee, Assaf Naor, and Yuval Peres. Trees and markov convexity, 2007.

[84] John M. Lee. *Introduction to Smooth Manifolds*. Springer, 2013.

[85] Matthias Leimeister and Benjamin J Wilson. Skip-gram word embeddings in hyperbolic space. *arXiv:1809.01498*, 2018.

[86] Yong-Lu Li, Xiaoqian Wu, Xinpeng Liu, Zehao Wang, Yiming Dou, Yikun Ji, Junyi Zhang, Yixing Li, Xudong Lu, Jingru Tan, et al. From isolated islands to pangea: Unifying semantic space for human action understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16582–16592, 2024.

[87] Zhen Li, Bing Xu, Conghui Zhu, and Tiejun Zhao. Clmlf: A contrastive learning and multi-layer fusion method for multimodal sentiment detection. *arXiv:2204.05515*, 2022.

[88] Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *NeurIPS*, 35:17612–17625, 2022.

[89] Yi-Lun Liao and Tess Smidt. Equiformer: Equivariant graph attention transformer for 3d atomistic graphs. In *ICLR*, 2023.

[90] Yi-Lun Liao, Brandon Wood, Abhishek Das*, and Tess Smidt*. EquiformerV2: Improved Equivariant Transformer for Scaling to Higher-Degree Representations. In *ICLR*, 2024.

[91] Qi Liu, Maximilian Nickel, and Douwe Kiela. Hyperbolic graph neural networks. In *NeurIPS*, pages 8230–8241, 2019.

[92] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv:1907.11692*, 2019.

[93] Qiyao Ma, Menglin Yang, Mingxuan Ju, Tong Zhao, Neil Shah, and Rex Ying. Harec: Hyperbolic graph-llm alignment for exploration and exploitation in recommender systems. *arXiv:2411.13865*, 2024.

[94] Jonathan Masci, Davide Boscaini, Michael Bronstein, and Pierre Vandergheynst. Geodesic convolutional neural networks on riemannian manifolds. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 37–45, 2015.

[95] Jiří Matoušek. On the distortion required for embedding finite metric spaces into normed spaces. *Israel Journal of Mathematics*, 93(1):333–344, 1996.

[96] Jiří Matoušek. On embedding trees into uniformly convex banach spaces. *Israel Journal of Mathematics*, 114(1):221–237, 1999.

[97] Jiří Matoušek. *Lectures on Discrete Geometry*. Springer, 2002.

[98] Hirotaka Matsumoto, Takahiro Mimori, and Tsukasa Fukunaga. Novel metric for hyperbolic phylogenetic tree embeddings. *Biology Methods and Protocols*, 6(1):bpab006, 2021.

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

[100] Pascal Mettes, Mina Ghadimi Atigh, Martin Keller-Ressel, Jeffrey Gu, and Serena Yeung. Hyperbolic deep learning in computer vision: A survey. *International Journal of Computer Vision*, pages 1–25, 2024.

[101] Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *EMNLP*, pages 2381–2391. Association for Computational Linguistics, 2018.

[102] Federico Monti, Davide Boscaini, Jonathan Masci, Emanuele Rodola, Jan Svoboda, and Michael M Bronstein. Geometric deep learning on graphs and manifolds using mixture model cnns. In *CVPR*, pages 5115–5124, 2017.

[103] John Nash. C^1 Isometric Embeddings. *Annals of Mathematics*, 60(3), 1954.

[104] Maximilian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. *NeurIPS*, 30, 2017.

[105] Maximillian Nickel and Douwe Kiela. Learning continuous hierarchies in the lorentz model of hyperbolic geometry. In *ICML*, pages 3779–3788, 2018.

[106] Yann Ollivier. Ricci curvature of markov chains on metric spaces. *Journal of Functional Analysis*, 256(3):810–864, 2009.

[107] Avik Pal, Max van Spengler, Guido Maria D'Amely di Melendugno, Alessandro Flaborea, Fabio Galasso, and Pascal Mettes. Compositional entailment learning for hyperbolic vision-language models. *arXiv:2410.06912*, 2024.

[108] Kiho Park, Yo Joong Choe, Yibo Jiang, and Victor Veitch. The geometry of categorical and hierarchical concepts in large language models. *arXiv:2406.01506*, 2024.

[109] Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *Forty-first ICML*, 2024.

[110] Arkil Patel, Satwik Bhattamishra, and Navin Goyal. Are NLP models really able to solve simple math word problems? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2080–2094. Association for Computational Linguistics, 2021.

[111] Aleksandar Poleksic. Hyperbolic matrix factorization improves prediction of drug-target associations. *Scientific Reports*, 13(1):959, 2023.

[112] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *arXiv*, 2021.

[113] 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.

[114] Sameera Ramasinghe, Violetta Shevchenko, Gil Avraham, and Ajanthan Thalaiyasingam. Accept the modality gap: An exploration in the hyperbolic space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27263–27272, 2024.

[115] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, et al. Zero-shot text-to-image generation. *ICML*, 2021.

[116] Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *KDD*, pages 3505–3506, 2020.

[117] P. L. Robinson. The sphere is not flat. In *The American Mathematical Monthly*, volume 113, 2006.

[118] Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. *arXiv:1907.10641*, 2019.

[119] Frederic Sala, Chris De Sa, Albert Gu, and Christopher Ré. Representation tradeoffs for hyperbolic embeddings. In *ICML*, pages 4460–4469. PMLR, 2018.

[120] Rik Sarkar. Low distortion delaunay embedding of trees in hyperbolic plane. In *International Symposium on Graph Drawing*, pages 355–366. Springer, 2011.

[121] Chandni Saxena, Mudit Chaudhary, and Helen Meng. Cross-lingual word embeddings in hyperbolic space. *arXiv:2205.01907*, 2022.

[122] Rui Shi, Wei Zeng, Zhengyu Su, Hanna Damasio, Zhonglin Lu, Yalin Wang, Shing-Tung Yau, and Xianfeng Gu. Hyperbolic harmonic mapping for constrained brain surface registration. In *Proceedings of the IEEE Conference on computer vision and pattern recognition*, pages 2531–2538, 2013.

[123] Ryohei Shimizu, Yusuke Mukuta, and Tatsuya Harada. Hyperbolic neural networks++. In *ICLR*, 2020.

[124] Jiajun Song and Yiqiao Zhong. Uncovering hidden geometry in transformers via disentangling position and context. *arXiv:2310.04861*, 2023.

[125] Mingyang Song, Yi Feng, and Liping Jing. Hisum: Hyperbolic interaction model for extractive multi-document summarization. In *Proceedings of the ACM Web Conference 2023*, pages 1427–1436, 2023.

[126] Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *arXiv preprint arXiv:2104.09864*, 2021.

[127] Li Sun, Zhongbao Zhang, Junda Ye, Hao Peng, Jiawei Zhang, Sen Su, and S Yu Philip. A self-supervised mixed-curvature graph neural network. In *AAAI*, volume 36, pages 4146–4155, 2022.

[128] Li Sun, Junda Ye, Hao Peng, Feiyang Wang, and S Yu Philip. Self-supervised continual graph learning in adaptive riemannian spaces. In *AAAI*, volume 37, pages 4633–4642, 2023.

[129] Li Sun, Zhenhao Huang, Zixi Wang, Feiyang Wang, Hao Peng, and Philip Yu. Motif-aware riemannian graph neural network with generative-contrastive learning. In *AAAI*, 2024.

[130] Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv:2406.06525*, 2024.

[131] Dídac Surís, Ruoshi Liu, and Carl Vondrick. Learning the predictability of the future. In *ICCV*, pages 12607–12617, 2021.

[132] Xunzhu Tang, Saad Ezzini, Haoye Tian, Yewei Song, Jacques Klein, Tegawende F Bissyande, et al. Hyperbolic code retrieval: a novel approach for efficient code search using hyperbolic space embeddings. *arXiv:2308.15234*, 2023.

[133] Yi Tay, Luu Anh Tuan, and Siu Cheung Hui. Hyperbolic representation learning for fast and efficient neural question answering. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 583–591, 2018.

[134] Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, et al. Gemma 2: Improving open language models at a practical size. *arXiv:2408.00118*, 2024.

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

[136] Alexandru Tifrea, Gary Bécigneul, and Octavian-Eugen Ganea. Poincaré glove: Hyperbolic word embeddings. In *ICLR*, 2019.

[137] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *arXiv:2302.13971*, 2023.

[138] Daniel J Trosten, Rwiddhi Chakraborty, Sigurd Løkse, Kristoffer Knutsen Wickstrøm, Robert Jenssen, and Michael C Kampffmeyer. Hubs and hyperspheres: Reducing hubness and improving transductive few-shot learning with hyperspherical embeddings. In *ICCV*, pages 7527–7536, 2023.

[139] Constantin Udriste. *Convex Functions and Optimization Methods on Riemannian Manifolds*, volume 297. Springer Science & Business Media, 1994.

[140] Eugenio Urdapilleta, Francesca Troiani, Federico Stella, and Alessandro Treves. Can rodents conceive hyperbolic spaces? *Journal of the Royal Society Interface*, 12(107):20141214, 2015.

[141] Max van Spengler, Erwin Berkhout, and Pascal Mettes. Poincaré resnet. *CVPR*, 2023.

[142] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, pages 5998–6008, 2017.

[143] Amelia Villegas-Morcillo, Victoria Sanchez, and Angel M Gomez. Foldhsphere: deep hyperspherical embeddings for protein fold recognition. *BMC bioinformatics*, 22:1–21, 2021.

[144] Hanzhang Wang, Jiawen Zhang, and Qingyuan Ma. Exploring intrinsic dimension for vision-language model pruning. In *Forty-first ICML*.

[145] Shen Wang, Xiaokai Wei, Cicero Nogueira Nogueira dos Santos, Zhiguo Wang, Ramesh Nallapati, Andrew Arnold, Bing Xiang, Philip S Yu, and Isabel F Cruz. Mixed-curvature multi-relational graph neural network for knowledge graph completion. In *TheWebConf*, pages 1761–1771, 2021.

[146] Maurice Weber, Daniel Y. Fu, Quentin Anthony, Yonatan Oren, Shane Adams, Anton Alexandrov, Xiaozhong Lyu, Huu Nguyen, Xiaozhe Yao, Virginia Adams, Ben Athiwaratkun, Rahul Chalamala, Kezhen Chen, Max Ryabinin, Tri Dao, Percy Liang, Christopher Ré, Irina Rish, and Ce Zhang. Redpajama: an open dataset for training large language models. *NeurIPS Datasets and Benchmarks Track*, 2024.

[147] Melanie Weber and Suvrit Sra. Projection-free nonconvex stochastic optimization on Riemannian manifolds. *IMA Journal of Numerical Analysis*, 42(4):3241–3271, 2021.

[148] Melanie Weber and Suvrit Sra. Riemannian Optimization via Frank-Wolfe Methods. *Mathematical Programming*, 2022.

[149] Melanie Weber, Manzil Zaheer, Ankit Singh Rawat, Aditya Menon, and Sanjiv Kumar. Robust large-margin learning in hyperbolic space. In *NeurIPS*, 2020.

[150] Yiwei Wei, Shaozu Yuan, Ruosong Yang, Lei Shen, Zhangmeizhi Li, Longbiao Wang, and Meng Chen. Tackling modality heterogeneity with multi-view calibration network for multimodal sentiment detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5240–5252, 2023.

[151] Richard C Wilson, Edwin R Hancock, Elżbieta Pekalska, and Robert PW Duin. Spherical and hyperbolic embeddings of data. *IEEE transactions on pattern analysis and machine intelligence*, 36(11):2255–2269, 2014.

[152] Marysia Winkels and Taco S Cohen. 3d g-cnns for pulmonary nodule detection. *arXiv:1804.04656*, 2018.

[153] Chenxiao Yang, Qitian Wu, and Junchi Yan. Geometric knowledge distillation: Topology compression for graph neural networks. In *NeurIPS*, volume 35, pages 29761–29775, 2022.

[154] Menglin Yang, Zhihao Li, Min Zhou, Jiahong Liu, and Irwin King. Hicf: Hyperbolic informative collaborative filtering. In *SIGKDD*, pages 2212–2221, 2022.

[155] Menglin Yang, Min Zhou, Jiahong Liu, Defu Lian, and Irwin King. Hrcf: Enhancing collaborative filtering via hyperbolic geometric regularization. In *Proceedings of the ACM Web Conference 2022*, pages 2462–2471, 2022.

[156] Menglin Yang, Min Zhou, Hui Xiong, and Irwin King. Hyperbolic temporal network embedding. *TKDE*, 35(11):11489–11502, 2022.

[157] Menglin Yang, Aosong Feng, Bo Xiong, Jihong Liu, Irwin King, and Rex Ying. Hyperbolic fine-tuning for large language models. *arXiv:2410.04010*, 2024.

[158] Menglin Yang, Aosong Feng, Bo Xiong, Jihong Liu, Irwin King, and Rex Ying. Hyperbolic fine-tuning for large language models. *ICML LLM Cognition Workshop*, 2024.

[159] Menglin Yang, Harshit Verma, Delvin Ce Zhang, Jiahong Liu, Irwin King, and Rex Ying. Hypformer: Exploring efficient transformer fully in hyperbolic space. In *KDD*, pages 3770–3781, 2024.

[160] Xin Yang, Heng Chang, Zhijian Lai, Jinze Yang, Xingrun Li, Yu Lu, Shuaiqiang Wang, Dawei Yin, and Erxue Min. Hyperbolic contrastive learning for cross-domain recommendation. In *CIKM*, page 2920–2929, 2024.

[161] Hongyi Zhang and Suvrit Sra. First-order methods for geodesically convex optimization. In *COLT*, 2016.

[162] Hongyi Zhang, Sashank J Reddi, and Suvrit Sra. Riemannian SVRG: Fast stochastic optimization on Riemannian manifolds. In *NeurIPS*, volume 29, 2016.

[163] Huanqiu Zhang, P Dylan Rich, Albert K Lee, and Tatyana O Sharpee. Hippocampal spatial representations exhibit a hyperbolic geometry that expands with experience. *Nature Neuroscience*, 26(1):131–139, 2023.

[164] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona T. Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. Opt: Open pre-trained transformer language models. *arXiv:2205.01068*, 2022.

[165] Yanqi Zhou, Tao Lei, Hanxiao Liu, Nan Du, Yanping Huang, Vincent Zhao, Andrew M Dai, Quoc V Le, James Laudon, et al. Mixture-of-experts with expert choice routing. In *NeurIPS*, volume 35, pages 7103–7114, 2022.

A Comprehensive Background and Related Works

A.1 Riemannian Geometry and Non-Euclidean Foundations

Riemannian Manifolds. A smooth n-dimensional manifold $\mathcal{M}$ is a topological space in which each point $x \in \mathcal{M}$ has a neighborhood $U_x \subseteq \mathcal{M}$ that is locally Euclidean, meaning that there exists a homeomorphism between U_x and an open subset of $\mathbb{R}^n$.

Tangent Space. Each point $x \in \mathcal{M}$ is associated with a *tangent space* $T_x\mathcal{M}$, which is an n-dimensional vector space serving as a first-order local approximation of $\mathcal{M}$ at x. This space encapsulates the possible directions in which one can move away from x on the manifold.

Riemannian Metric. A *Riemannian metric* $\mathfrak{g}$ on $\mathcal{M}$ is a collection of positive-definite bilinear forms $\mathfrak{g}_x(\cdot,\cdot) : T_x\mathcal{M} \times T_x\mathcal{M} \to \mathbb{R}$, smoothly varying with x. The metric $\mathfrak{g}_x$ induces the *sectional curvature* at each point, which measures the extent to which the manifold deviates from flatness at x. A *Riemannian manifold* is then defined as the pair $(\mathcal{M}, \mathfrak{g})$. For instance, $\mathbb{R}^n$ with the usual Euclidean inner product is a Riemannian manifold with zero curvature. The metric $\mathfrak{g}_x$ generalizes the notion of inner products, with the norm of a vector $p \in T_x\mathcal{M}$ given by $\|p\|_{\mathfrak{g}} = \sqrt{\mathfrak{g}_x(p,p)}$. The choice of the Riemannian metric also induces a global distance function $d(\cdot,\cdot)$ on $\mathcal{M}$.

Geodesic. A *geodesic* between two points x and y is a smooth curve that locally minimizes the distance between these points. In particular, the shortest path between x and y is a geodesic.

Exponential Map. Under certain conditions, one can define the *exponential map* $\exp_x : T_x\mathcal{M} \to \mathcal{M}$, which lifts points from the tangent space $T_x\mathcal{M}$ to the manifold $\mathcal{M}$, by associating a vector in $T_x\mathcal{M}$ to a point on $\mathcal{M}$ along a geodesic.

Logarithmic Map. The *logarithmic map* $\log_x : \mathcal{M} \to T_x\mathcal{M}$ is the inverse of the exponential map, provided certain assumptions on $\mathcal{M}$ hold.

Geodesics and Geodesic Operations. The Riemannian metric $\mathfrak{g}_x$ can be viewed as a generalization of the inner product, where the norm of a vector $p \in T_x\mathcal{M}$ is defined by $\|p\|_{\mathfrak{g}} = \sqrt{\mathfrak{g}_x(p,p)}$. The choice of $\mathfrak{g}$ induces a global distance function $d(\cdot,\cdot)$ on $\mathcal{M}$, where geodesics are the locally distance-minimizing curves. The length of a geodesic between two points determines the geodesic distance. The exponential map $\exp_x$ maps a vector $v \in T_x\mathcal{M}$ to a point on $\mathcal{M}$ along the geodesic starting at x. The logarithmic map $\log_x$ is the inverse of this process. Additionally, the parallel transport map $\mathrm{PT}_x(v,w)$ transports vectors along geodesics, providing a canonical way to move vectors in a manner consistent with the underlying geometric structure. It canonically transports a vector w along a geodesic emanating from x with initial velocity v and zero acceleration. This generalizes the classical notion of translation in Euclidean space.

Hyperbolic Spaces. Hyperbolic spaces are Riemannian manifolds with constant negative curvature, i.e., with curvature $-K < 0$. Common models for hyperbolic space include the *Poincaré ball model* $\mathbb{P}^{K,n}$ and the *Lorentz hyperboloid* $\mathbb{L}^{K,n}$, which have been extensively studied in the context of deep learning [104, 47]. For points $\mathbf{x}, \mathbf{y} \in \mathbb{L}^{K,n}$, their inner product $\langle \mathbf{x}, \mathbf{y} \rangle_{\mathcal{L}}$ is given by $\langle \mathbf{x}, \mathbf{y} \rangle_{\mathcal{L}} = -x_t y_t + \mathbf{x}_s^T \mathbf{y}_s = \mathbf{x}^T \mathfrak{g}_n^K \mathbf{y}$ with $|||\mathbf{x}|||_{\mathcal{L}} \coloneqq \sqrt{|\langle \mathbf{x}, \mathbf{x} \rangle_{\mathcal{L}}|}$ being the Lorentzian norm. Formally, $\mathcal{L}^n$ is the point set $\mathcal{L}^n = \{\mathbf{x} \in \mathbb{R}^{n+1} : \langle \mathbf{x}, \mathbf{x} \rangle_{\mathcal{L}} = 1/K, x_t > 0\}$. $\mathbb{P}^{n,K}$ is the n-dimensional sphere S^n with radius $1/\sqrt{K}$, with the Riemannian metric $g_x^{\mathbb{P}} = \lambda_x^2 g^E$, where $\lambda_x := \frac{2}{1-c\|x\|^2}$ and $\mathfrak{g}^E$ is the Euclidean metric. Other models, such as the Klein model, also exist. These models are *isometric*, meaning that there is a smooth correspondence between points in different models that preserves distances, angles, and geodesics. This property allows for the selection of the most suitable model for a given application.

Spherical Spaces. Spherical spaces are Riemannian manifolds with constant positive curvature, i.e., with curvature $K > 0$. An n-dimensional spherical space $\mathbb{S}^{K,n}$ is an n-dimensional sphere of radius $K^{-\frac{1}{2}}$, equipped with the Riemannian metric induced by the Euclidean metric on $\mathbb{R}^{n+1}$.

Mixed Curvature Spaces. A *mixed curvature space* $\mathcal{M}$ is defined as a product manifold consisting of Euclidean, spherical, and hyperbolic spaces. The Riemannian metric and geodesic operations for such a manifold are defined component-wise, enabling effective computational implementation for downstream tasks.

Generalized Riemannian Manifolds. Generalizations of Riemannian manifolds can be obtained by relaxing some of the assumptions in their classical definition. One notable generalization is the *pseudo-Riemannian manifold*, in which the metric $\mathfrak{g}$ is an indefinite bilinear form, allowing for both positive and negative signs. This generalization is useful in contexts such as relativistic physics, where spacetime is modeled as a pseudo-Riemannian manifold.

A.2 Non-Applicability of the Nash Embedding Theorem

The Nash Embedding Theorem roughly states that any Riemannian manifold of dimension n admits an isometric embedding into $\mathbb{R}^{2n+1}$ [103]. While it may appear as if this allows for Euclidean embeddings of complex structures with no distortion and only twice the dimension, this is in fact a confusion in vocabulary between the notion of isometric embeddings between those of *Riemannian manifolds* and those of *metric spaces*.

Definition A.1. Let $(\mathcal{M}, \mathfrak{g}), (\mathcal{M}', \mathfrak{g}')$ be Riemannian manifolds. An *isometric embedding of Riemannian manifolds* is a smooth map $f : \mathcal{M} \to \mathcal{M}'$ such that $\mathfrak{g} = f^*\mathfrak{g}'$. Let $(X, d_X), (Y, d_Y)$ be metrics spaces. An *isometric embedding of metric spaces* is a map $f : X \to Y$ such that $d_X(a, b) = d_Y(f(a), f(b))$ for all $a, b \in X$.

Hence in the former, which is also the isometric embedding afforded by the Nash Embedding Theorem, the map f preserves the Riemannian metric, i.e. the inner product on the tangent bundle. As a result, the isometry is *locally distance preserving*, in the sense that *length of any path* is preserved under f. However, given points x, y connected by a shortest path γ, the straight line path connecting $f(x), f(y)$ in the co-domain is not necessarily $f(\gamma)$ (note that f need not to be surjective). As a result, measuring the distortion of embeddings is concerned with whether f is *globally distance preserving*, or whether the shortest distance between $f(x)$ and $f(y)$ is the length of $f(\gamma)$, which is defined by isometric embeddings between metric spaces. Note that **an isometric embedding of Riemannian manifolds is in general not an isometric embedding of metric spaces**. For instance, given the sphere S^1, its usual Riemannian metric is inherited from the Riemannian metric for $\mathbb{R}^2$, i.e. the usual inner product. The identity map is then an isometric embedding $S^1 \hookrightarrow \mathbb{R}^2$ as Riemannian manifolds. However, the distance between points on the sphere does not coincide with the Euclidean distance of their image. As an example, antipodal points have distance π in S^1 but distance 1 in $\mathbb{R}^2$.

In the context of foundational models, we are concerned with globally distance preserving embeddings, as computing the distance between any pairs of token embeddings is crucial for model training. As a result, the Nash Embedding Theorem is not applicable since global distortion could still arise from isometries between Riemannian manifolds. For this reason, by "isometry", we refer to those between metric spaces unless otherwise specified, which captures the notion of distortion critical for foundational model embeddings.

Definition A.2. Let $(X, d_X), (Y, d_Y)$ be metric spaces equipped with the respective distance metrics and $f : X \to Y$ be a map. The *bi-Lipschitz distortion* of f is $\mathrm{dist}(f) = \|f\|_{\mathrm{Lip}}\|f^{-1}\|_{\mathrm{Lip}}$, where $\|f\|_{\mathrm{Lip}}$ is the (possibly infinite) Lipschitz-constant of f. For a pair of points $(a, b) \in X^2$, the *point-wise distortion* is given by $\frac{|d_X(a,b) - d_Y(f(a)-f(b))|}{d_X(a,b)}$.

Both notions of distortion measure the deviation of f from an *isometry between metric spaces*. Note that the minimum distortion in the case of bi-Lipschitz distortion is 1.

A.3 Markov Convexity

In this section we provide the relevant background on notion of Markov convexity. Let (X, d_X) be a metric space. Then the *Markov p–convexity constant* Π (for a fixed positive integer p) of the metric space X is a universal constant (or ∞) define as follows:

Definition A.3. Π is the smallest constant s.t. for any Markov chain on $(X_t)_{t\geq 0}$ on a state space Ω, and every map $f : \Omega \to X$, and for any $m \in \mathbb{N}$, we have

$$\sum_{n=0}^{\infty} \frac{1}{2^{np}} \sum_{t\in\mathbb{Z}} \mathbb{E}\Big[d\big(f(X_t), f(X_{t+2^n})\big)^p\Big] \;\leq\; \Pi^p \sum_{t\in\mathbb{Z}} \mathbb{E}\Big[d\big(f(X_t), f(X_{t+1})\big)^p\Big]$$

Roughly speaking, when $\Pi < \infty$, the p-th moment of one-step increments dominates the p-th moments of exponential length steps. Intuitively, measures how tightly local behaviors in X control and estimate global behaviors on the space, with lower values Π showing tighter control.

A.4 Non-Euclidean Structure in the Real World

Non-Euclidean Structures in Natural Language Processing. Language exhibits inherently hierarchical structures - from concept taxonomies to entailment relationships - that challenge traditional Euclidean representations. These hierarchical relationships between linguistic units naturally manifest on non-Euclidean manifolds, particularly in hyperbolic space, which has emerged as a powerful framework for natural language processing [37, 85]. Foundational work has demonstrated that hyperbolic embeddings can effectively capture word-level semantics [136] and concept hierarchies [82], leveraging the exponential volume growth of hyperbolic space to model tree-like linguistic structures. The success of hyperbolic representations has sparked various advanced applications: from question answering systems [133], privacy-preserving text representations [45], to multi-document summarization that captures document-level discourse structure [125]. Recent work has further extended these approaches to cross-lingual settings [121] and contextual language models [24], demonstrating the broad utility of non-Euclidean geometries in modern natural language processing.

Non-Euclidean Structures in Computer Vision. Similar to NLP, many computer vision tasks involve data that naturally resides in intricate manifolds that are challenging to model using conventional Euclidean space [100]. For instance, visual entities often form inherent hierarchical relationships among object classes, between scenes and their constituent categories [51, 107], or scenes at varying levels of granularity [73]. In these scenarios, hyperbolic geometry provides a compelling alternative to the Euclidean representations in representing the exponential growth of hierarchical structures with minimal distortion [119]. Its advantages have been demonstrated across a wide range of applications, including image segmentation [4], action classification [25] video prediction [131], deformable 3D surfaces [94]. In parallel, hyperspherical learning has become integral to modern contrastive learning with cosine similarity, underpining tasks ranging from self-supervised learning [40] to long-tailed classification [71] and few-shot learning [138].

Non-Euclidean Structures in Complex Networks. Networks, whether they represent social interactions, user purchasing preferences, or transportation systems, often exhibit complex, non-Euclidean relationships that traditional Euclidean models fail to capture effectively. Social networks, for example, are best described by graph structures where nodes (individuals) are connected by edges (relationships) that can be directional, weighted, or even exhibit hierarchical properties. These networks typically involve intricate dependencies and nonlinear relationships, requiring geometric frameworks beyond Euclidean space to model effectively.

Non-Euclidean Structures in Natural Sciences. In natural science, many systems exhibit intricate structures that Euclidean space struggles to capture effectively. In biology, non-Euclidean geometries are integral to analyzing and modeling complex organic structures, such as protein folding [143], single-cell RNA-seq data [76, 38, 13], and phylogenetic trees [98], where hyperbolic and spherical geometries are commonly observed. In neuroscience, hyperbolic geometry is shown to be more effective than Euclidean counterpart in modeling the brain's cortical folding [140], brain surface [122], and hippocampal spatial representations [163], aiding in the study of spatial organization and connectivity.

A.5 Deep Learning with Non-Euclidean Geometries

Recent years have witnessed an increasing interest in extending deep learning techniques to Riemannian manifolds. Here we discuss in further detail the advances for designing neural networks and Transformers in non-Euclidean geometries, as well as optimization on manifolds.

Geodesic Neural Networks. Geodesic neural networks leverage geodesic, particularly geodesic distances, to perform neural operations that preserve geometric structure on manifold-structured data [21]. Several works have developed geodesic convolutional layers by applying filters to local patches in geodesic polar coordinates [94], learning directionally sensitive filters along principal curvature directions [17], or learnable kernel functions that operate on local coordinate systems [102]. More recent works such as GDGNN [78] have integrated geodesic operations with graph representations.

Hyperbolic Neural Networks. Hyperbolic neural networks exploit the geometry of hyperbolic space to learn embeddings that reflect hierarchical relationships more effectively than their Euclidean counterparts [104]. HNN [47] and HNN++[123] developed many basic operations, such as hyperbolic linear and convolutional layers, and multinomial logistic regression (MLR). HGCN [23] and HGNN [91] were then among the first to develop hyperbolic graph neural networks (GNNs). More recently, HyboNet [26] proposed a framework of hyperbolic neural networks that does not depend on the Euclidean tangent spaces; Poincaré ResNet [141] and HCNN [10] developed components for hyperbolic vision models; LResNet [63] proposed an efficient and stable residual connection method.

Spherical Neural Networks. Spherical neural networks are designed for data that naturally reside on spheres or benefit from spherical symmetry. Spherical CNNs [31, 41] extended convolutions and pooling to preserve rotational symmetries. SphereNet [32] introduced a framework for learning spherical image representations by encoding distortion invariance into convolutional filters. DeepSphere [34] proposed a graph-based approach. SWSCNN [43] later proposed a fully spherical CNN that allows for anisotropic filters.

Mix-curvature Neural Networks. Mix-curvature neural networks uses product spaces of the aforementioned manifolds to better model data that have local neighborhoods exhibiting different geometric properties. [55] developed the first learning framework on product spaces, introducing fundamental techniques such as mean and loss functions for embedding optimization. κ-GCN [6] then extended learning on product spaces to GCNs, introducing a unified and differentiable Gyrovector spaces framework to constant curvature spaces beyond hyperbolic manifolds.

Non-Euclidean Transformers. Significant advancements have been made toward Transformers in non-Euclidean spaces in recent studies. Within hyperbolic learning, several works have proposed hyperbolic self-attention mechanisms [56, 26, 123] and hyperbolic linear attentions [159], enabling constructions of hyperbolic Transformers. Hyperbolic fine-tuning methods have also been developed for LLMs [158]. Recent works have also proposed hyperbolic vision Transformers [44]. Attention mechanisms have been developed for spherical spaces as well [81]. Further, Transformers have been developed for mixed curvature manifolds as well [29].

Optimization on manifolds. Learning on manifolds often times require optimizing parameters with manifold constraints. Many classical convex optimization algorithms have been extended to the manifold-valued setting [139, 5, 161, 148]. Stochastic optimization on manifolds has been studied extensivley [16, 162, 11, 147], which includes extensions of algorithms such as SGD and Adam, which are suitable for training models on geometric domains.

B Additional Statistics and Dataset Details

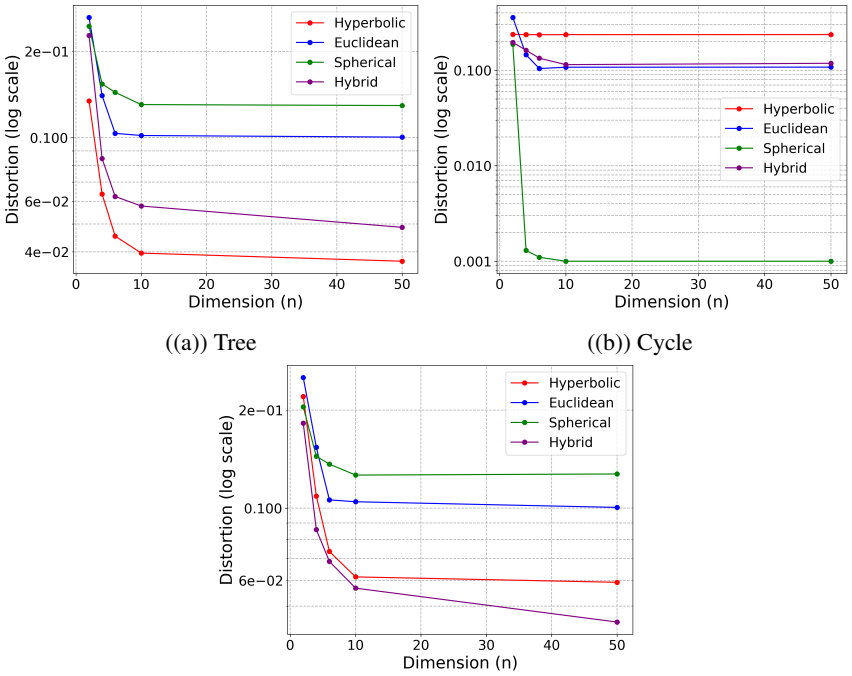
In this section we give details regarding the datasets we used, as well as the show more statistic results for more LLMs. We also show the distortion v.s. dimensionality plot for all graph here.

B.1 Distortion v.s. Dimensionality

In this section we provide more plots of the distortion of embedding graphs into manifold of varying dimensions. The plots are shown in Figure 6. In all cases, non-Euclidean geometry achieves significantly smaller distortion with significantly fewer dimensions, reflecting Takeaway 1 in Section 3.1. The distortion for Euclidean embeddings always plateaus, demonstrating that it is not suited for embeddings each structures regardless of its dimension. On the other hand, the distortion for non-Euclidean embeddings is still being reduced with increased dimensionality for 2 of the structures. This reflects Takeaway 2 in Section 3.1.

B.2 Dataset Details

For the evaluation of token embedding distribution in LLMs, we incorporated a wide range of datasets, including a subset of the RedPajama dataset [146] encompassing the arXiv, C4, Common Crawl, GitHub, Wikipedia, and StackExchange datasets; math reasoning datasets such as GSM8K [30], MATH50K [64],MAWPS [77], and SVAMP [110]; and common sense reasoning datasets, including BoolQ, WinoGrande [118], and OpenBookQA [101].



((a)) Tree

((b)) Cycle

((c)) Ring of Trees

Figure 6: Distortion of embedding a complete tree, cycle, and ring of tree into manifolds of different dimensions (log scale). Each graph has 96 nodes. Euclidean embeddings is shown in blue. In all cases, non-Euclidean geometry achieves significantly smaller distortion with significantly fewer dimensions. The distortion for Euclidean embeddings always plateaus, demonstrating that it is not suited for embeddings each structures regardless of its dimension.

Table 2: Hyperbolicity values δ for different metric spaces.

	Sphere Space	Dense Graph	PubMed Graph	Poincare Space	Tree Graph
δ	0.99 ± 0.01	0.62 ± 0.01	0.40 ± 0.04	0.14 ± 0.01	0.0

B.3 More Statistics

In Figure 7 we show the statistics for token embeddings for more LLMs, including GPT-NeoX-20B [14], OPT-13B [164], RoBERT-Base [92], Gemma2-9B [134], LLaMa3.1-8B [53], and LLaMa-13B [137]. The top 2 rows show distribution of the norm of the token embeddings and the bottom 2 rows show the distribution of the frequency of each token embedding. The token frequency distribution demonstrate scale-free property with power law decay, whereas the token norm show rapid decreases in token count for higher normed tokens at the right tail. However, still none of the Euclidean foundational models fully capture the underlying scale-free property of the distribution, with all of them having an initial increase in token count against token norm for small normed token embeddings.

C δ-Hyperbolicity Computation

Given any four points a, b, c, and w in a metric space, the Gromov product $[a, c]_w$ at w is bounded below by the minimum of the Gromov products $[a, b]_w$ and $[b, c]_w$, minus a slack term δ:

$$[a, c]_w \geq \min([a, b]_w, [b, c]_w) - \delta. \tag{1}$$

The Gromov product between a and b with respect to w is defined as:

$$[a, b]_w = \frac{1}{2} \left(d(a, w) + d(b, w) - d(a, b) \right). \tag{2}$$

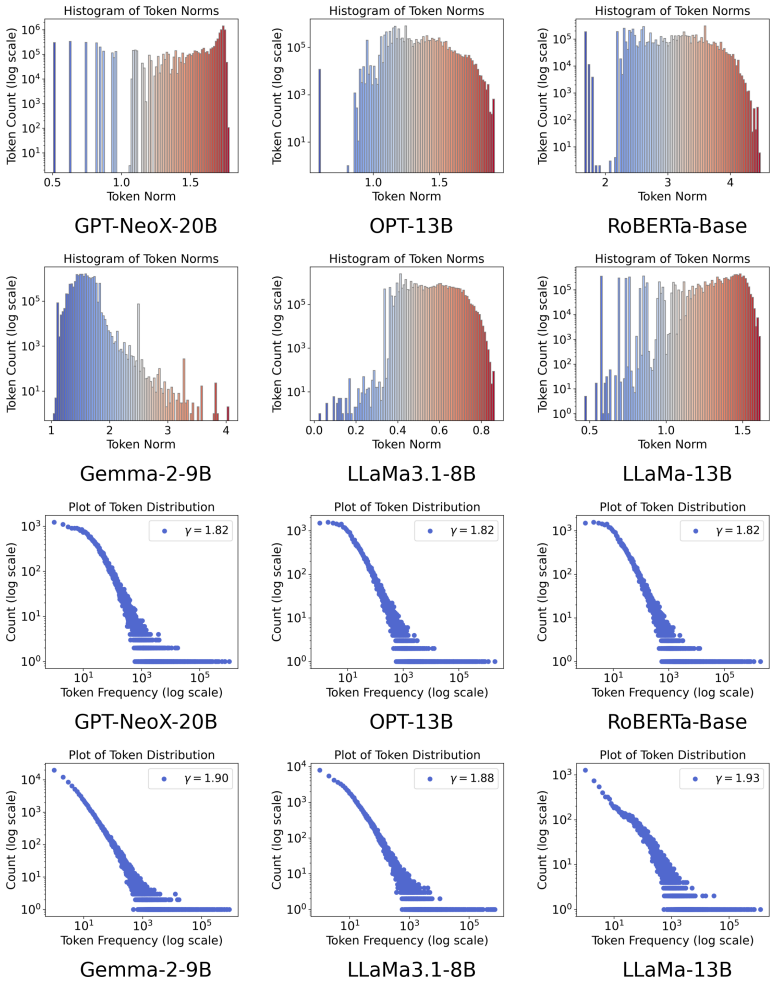


Figure 7: Token embeddings statistics for GPT-NeoX-20B, OPT-13B, RoBERT-Base, Gemma2-9B, LLaMa3.1-8B, and LLaMa-13B. The top 2 rows show distribution of the norm of the token embeddings and the bottom 2 rows show the distribution of the frequency of each token embedding. The token frequency distribution demonstrate scale-free property with power law decay, whereas the token norm show rapid decreases in token count for higher normed tokens at the right tail.

.

A metric space X is said to be δ-hyperbolic if this inequality holds for all choices of a, b, c, and w. In geodesic metric spaces, δ-hyperbolicity implies that geodesic triangles satisfy the δ-slim property, meaning that any point on one side of a geodesic triangle is at most a distance of δ from some point on one of the other two sides.

In an exact tree metric, where the sides of any triangle intersect at a single point, the hyperbolicity constant δ is zero. This follows from the fact that the four-point condition holds as an equality for all points in the space.

D Foundational Operations for Pretraining Non-Euclidean Foundation Models

Table 3 systematically compares foundational operations in Euclidean space with their adaptations to non-Euclidean manifold spaces, highlighting critical geometric modifications required for pretraining curvature-aware foundation models. Below, we explain the key components and their mathematical formulations:

Table 3: Geometric Foundation Model Operations: Euclidean vs. Manifold Formulations. $\mathrm{PT}_{\mathcal{M}}$:Parallel transport preserving vector properties during translation; $\exp_{\mu_{\mathcal{M}}}$: Exponential map from tangent space at Fréchet mean $\mu_{\mathcal{M}}$; $\log_{\mu_{\mathcal{M}}}$: Inverse exponential map projecting to tangent space; Ret: Retraction mapping for parameter updates; Proj: Tangent space projection operator

Operation	**Euclidean Space**	**Manifold Space**
Curvature (K)	$K = 0$	$K \in \mathbb{R}$
Attention Score	$\alpha_{qk} = \text{softmax}\left(\frac{q \cdot k^\top}{\sqrt{d_k}}\right)$	$\alpha_{qk}^{\mathcal{M}} = \text{softmax}\left(\frac{-d^2_{\mathcal{M}}(q,k)}{\sqrt{d_k}}\right)$
Rotary PE	$Q_i^{\text{RoPE}} = Q_i \,\mathrm{Rot}(\mathbf{p}_i); \quad K_i^{\text{RoPE}} = K_i \,\mathrm{Rot}(\mathbf{p}_i)$	$Q_i^{\text{RoPE}_{\mathcal{M}}} = \mathrm{PT}_{\mathcal{M}}(Q_i, \mathbf{p}_i); \quad K_i^{\text{RoPE}_{\mathcal{M}}} = \mathrm{PT}_{\mathcal{M}}(K_i, \mathbf{p}_i)$
Residual Connection	$x^{(l+1)} = x^{(l)} + f(x^{(l)})$	$x^{(l+1)} = \exp_{x^{(l)}}(\lambda \cdot f(x^{(l)}))$
Layer Norm	$\hat{x} = \frac{x-\mu}{\sigma}$	$\hat{x} = \exp_{\mu_{\mathcal{M}}}\left(\frac{\log_{\mu_{\mathcal{M}}}(x)}{\sigma_{\mathcal{M}}}\right)$
Cross-Entropy Loss	$\mathcal{L} = -\sum_t \log p_t$	$\mathcal{L} = -\sum_t \log \frac{\exp(-d_{\mathcal{M}}(z_t, z^*))}{\sum_{t'} \exp(-d^2_{\mathcal{M}}(z_t, z_{t'}))}$
Optimization	$\theta_{t+1} = \theta_t - \eta \nabla_\theta J(\theta)$	$\theta_{t+1} = \mathrm{Ret}_{\theta_t}\left(-\eta \mathrm{Proj}_{T_{\theta_t}\mathcal{M}} \nabla J\right)$
FFN	$y = W_2\sigma(W_1 x + b_1) + b_2$	$y = \exp_{\mathbf{0}}\left(W_2\sigma(\log_{\mathbf{0}}(W_1 \otimes x \oplus b_1))\right)$
Attention Aggregation	$h = \sum_i \alpha_i v_i$	$h = \mathrm{WeightedExpSum}(\{v_i\}, \{\alpha_i\})$

Curvature (K). In Euclidean space, curvature is fixed at $K = 0$, reflecting flat geometry. In manifold spaces, curvature $K \in \mathbb{R}$ is a learnable or estimated parameter that defines the intrinsic geometry (hyperbolic $K < 0$, spherical $K > 0$, or mixed). This value influences all subsequent operations, requiring dynamic adjustments to distance metrics and parameter updates. Curvature estimation methods (e.g., Ollivier-Ricci for graphs or learned embeddings for non-graph data) ensure geometric consistency across tasks.

Attention Mechanism. Euclidean attention computes similarity via dot products $\alpha_{qk} = \text{softmax}\left(\frac{q \cdot k^\top}{\sqrt{d_k}}\right)$, while manifold attention replaces this with geodesic distance: $\alpha_{qk}^{\mathcal{M}} = \text{softmax}\left(\frac{-d^2_{\mathcal{M}}(q,k)}{\sqrt{d_k}}\right)$. The negative squared distance prioritizes proximity on the manifold, preserving geometric relevance. Aggregation uses weighted Fréchet means (via exponential maps) or tangent space projections to combine features without violating curvature constraints.

Positional Encoding (Rotary PE). Euclidean positional encodings apply rotation matrices $\mathrm{Rot}(\mathbf{p}i)$ to query/key vectors. For manifolds, parallel transport $\mathrm{PT}_{\mathcal{M}}$ replaces rotations, translating positional shifts along geodesics while preserving vector orientation relative to the manifold's curvature. This ensures positional relationships respect intrinsic geometry.

Residual Connections. Standard residuals $x^{(l+1)} = x^{(l)} + f(x^{(l)})$ are replaced by manifold equivalents: $x^{(l+1)} = \exp_{x^{(l)}}(\lambda \cdot f(x^{(l)}))$. Here, the exponential map $\exp$ projects tangent space updates $f(x^{(l)})$ onto the manifold, scaled by λ, to preserve geometric stability across layers.

Layer Normalization. Euclidean layer norm standardizes features via $\hat{x} = \frac{x-\mu}{\sigma}$. On manifolds, operations occur in the tangent space at the Fréchet mean $\mu_{\mathcal{M}}$: $\hat{x} = \exp_{\mu_{\mathcal{M}}}\left(\frac{\log_{\mu_{\mathcal{M}}}(x)}{\sigma_{\mathcal{M}}}\right)$, where $\log_{\mu_{\mathcal{M}}}$ maps points to the tangent space for normalization before reprojection.

Cross-Entropy Loss. The manifold loss $\mathcal{L} = -\sum_t \log \frac{\exp(-d_{\mathcal{M}}(z_t, z^*))}{\sum_{t'} \exp(-d^2_{\mathcal{M}}(z_t, z_{t'}))}$ replaces Euclidean dot products with geodesic distances, ensuring probabilities reflect the manifold's geometry. This penalizes deviations in the curved space rather than in a flat embedding.

Optimization. Euclidean SGD $\theta_{t+1} = \theta_t - \eta \nabla_\theta J(\theta)$ is adapted via retractions Ret_{θ_t}, which map gradient steps $-\eta \mathrm{Proj}_{T_{\theta_t}\mathcal{M}} \nabla J$ from the tangent space back to the manifold, ensuring updates respect curvature constraints.

Feed-Forward Network (FFN). Manifold FFNs $y = \exp_{\mathbf{0}}\left(W_2\sigma(\log_{\mathbf{0}}(W_1 \otimes x \oplus b_1))\right)$ use Möbius operations $(\otimes, \oplus)$ for linear transformations and biases, followed by activation in the tangent space. The exponential map $\exp_{\mathbf{0}}$ ensures outputs remain on the manifold.

Attention Aggregation. Instead of weighted sums $h = \sum_i \alpha_i v_i$, manifolds use $\mathrm{WeightedExpSum}$, which computes Fréchet means of values v_i weighted by α_i, ensuring aggregated features lie on the manifold.

These adaptations collectively enable pretraining in non-Euclidean spaces by preserving geometric integrity. Operations like parallel transport, exponential/log maps, and retractions ensure compatibility with curvature, while specialized normalization and loss functions align learning dynamics with the manifold's intrinsic structure. The table underscores the necessity of redefining core components—from attention to optimization—to build effective foundation models for hyperbolic and mixed-curvature geometries.