

## Appendix Contents

<b>A</b>	<b>More Investigation Results</b>	<b>25</b>
A.1	Token Frequency and Norm Distribution on Mathematical Reasoning . . . . .	25
A.2	Token Frequency and Norm Distribution on Commonsense Reasoning . . . . .	25
A.3	Hyperbolicity in the Final Hidden Layer of LLMs . . . . .	27
<b>B</b>	<b>Hyperbolicity on Different Metric Spaces</b>	<b>28</b>
<b>C</b>	<b>Exponential and Logarithmic Maps</b>	<b>28</b>
C.1	Exponential Map . . . . .	28
C.2	Logarithmic Map . . . . .	29
C.3	Notation in the Main Text . . . . .	29
<b>D</b>	<b>Lorentz Transformation</b>	<b>29</b>
D.1	Lorentz Boost . . . . .	29
D.2	Lorentz Rotation . . . . .	29
<b>E</b>	<b>Transformation Analysis</b>	<b>30</b>
<b>F</b>	<b>Full Comparison</b>	<b>32</b>
F.1	Implementation Details . . . . .	32
F.2	Comparison on Mathematical Reasoning . . . . .	33
F.3	Comparison on Commonsense Reasoning . . . . .	34
F.4	GPU Usage . . . . .	34
<b>G</b>	<b>Case Study</b>	<b>35</b>

## A More Investigation Results

### A.1 Token Frequency and Norm Distribution on Mathematical Reasoning

To provide a comprehensive understanding of the geometric properties of token embeddings across different mathematical reasoning tasks, we extend our analysis beyond the GSM8K dataset presented in the main text to include AQuA and MAWPS datasets. This broader investigation allows us to validate the consistency of our findings across diverse mathematical problem types and complexity levels. The AQuA dataset presents algebraic word problems that require multi-step reasoning and equation solving, while MAWPS focuses on elementary arithmetic word problems with varying structural complexity. By analyzing token distributions across these complementary datasets, we can assess whether the observed power-law behavior and hierarchical token organization represent universal properties of mathematical reasoning tasks or are specific to particular problem domains.

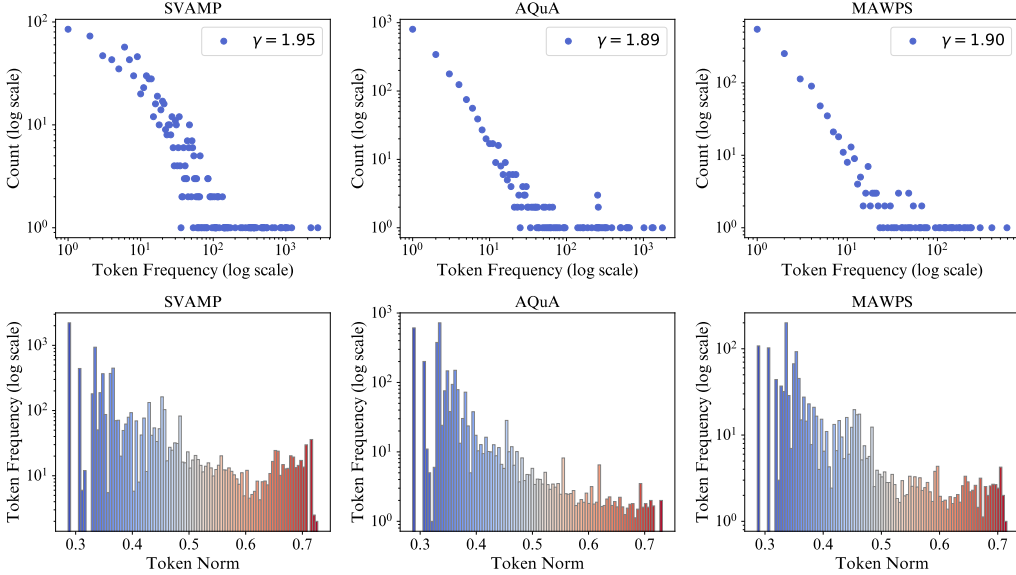


Figure 3: Token frequency distribution (top row) and token frequency vs. norm (bottom row) across different mathematical reasoning datasets in LLaMA3. The top row shows the power-law distribution of token frequencies with the decay rate ( $\gamma$ ) annotated for each dataset. The bottom row illustrates the relationship between token frequency and token norm, binned and colored by frequency, where higher token norms correspond to lower frequencies.

Our extended analysis, illustrated in Figure 3, reveals remarkably consistent patterns across all three mathematical reasoning datasets. The power-law exponents remain stable within a narrow range ( $\gamma \in [1.89, 1.95]$ ), indicating that the hierarchical structure of mathematical language is preserved regardless of the specific problem type or complexity level. The relationship between token frequency and embedding norms shows consistent inverse correlation across all datasets, with high-frequency mathematical operators and common function words clustering near the origin, while domain-specific mathematical terms and numerical values are positioned at greater distances. **This consistency strengthens our hypothesis that mathematical reasoning tasks inherently exhibit hyperbolic characteristics in their token embedding spaces**, providing strong empirical support for the effectiveness of hyperbolic fine-tuning approaches like HypLoRA in mathematical domains.

### A.2 Token Frequency and Norm Distribution on Commonsense Reasoning

To demonstrate the generalizability of our findings beyond mathematical reasoning, we conduct a comprehensive analysis of token distributions across six diverse commonsense reasoning datasets: ARC-Challenge, ARC-Easy, BoolQ, HellaSwag, PIQA, and SIQA. These datasets span a wide range of commonsense reasoning tasks, from factual knowledge retrieval (ARC datasets) and yes/no question answering (BoolQ) to physical commonsense (PIQA) and social understanding (SIQA).

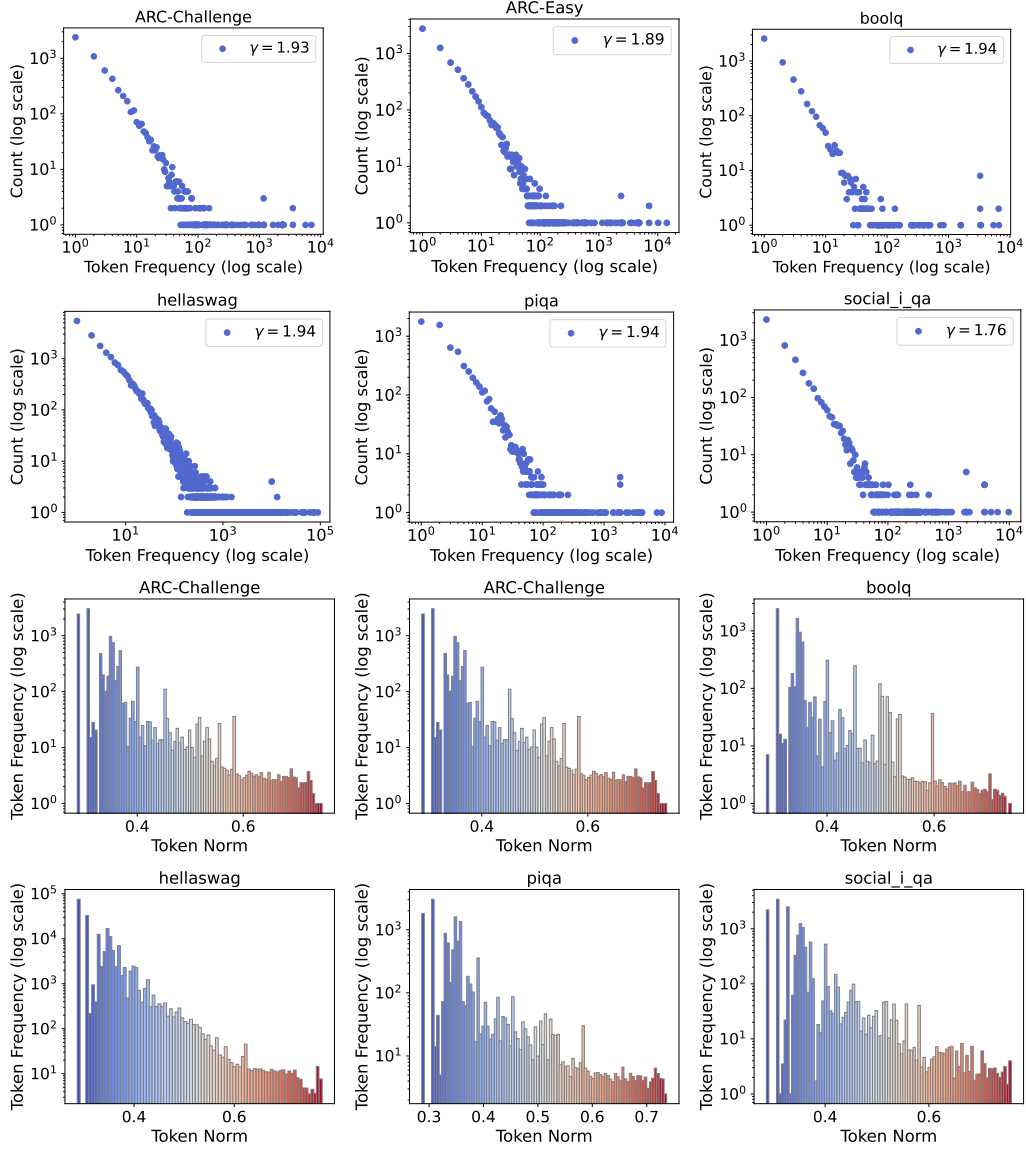


Figure 4: Token frequency distribution (top two rows) and token frequency vs. norm (bottom two rows) across different commonsense reasoning datasets in LLaMA3. The top two rows show the power-law distribution of token frequencies with the decay rate ( $\gamma$ ) annotated for each dataset. The bottom two rows illustrate the relationship between token frequency and token norm, binned and colored by frequency, where higher token norms correspond to lower frequencies.

Table 6: Relative  $\delta$ -hyperbolicity (mean  $\pm$  std.) of the final hidden layer in Gemma-7B across math (AQuA, GSM8K) and commonsense (ARC-Challenge, WinoGrande, OpenBookQA) datasets comparing the frozen base model, LoRA, DoRA, and HypLoRA.

Dataset	Base Model	LoRA	DoRA	HypLoRA
AQuA	$0.31 \pm 0.04$	$0.24 \pm 0.05$	$0.23 \pm 0.05$	$0.22 \pm 0.03$
GSM8K	$0.28 \pm 0.04$	$0.21 \pm 0.05$	$0.21 \pm 0.05$	$0.20 \pm 0.03$
ARC-Challenge	$0.30 \pm 0.03$	$0.35 \pm 0.03$	$0.36 \pm 0.02$	$0.25 \pm 0.02$
Winogrande	$0.22 \pm 0.04$	$0.32 \pm 0.02$	$0.27 \pm 0.02$	$0.27 \pm 0.02$
OpenbookQA	$0.30 \pm 0.03$	$0.35 \pm 0.03$	$0.38 \pm 0.02$	$0.25 \pm 0.02$

Table 7: Relative  $\delta$ -hyperbolicity (mean  $\pm$  std.) of the final hidden layer in Gemma3-4B for the same five datasets, contrasting the base model with LoRA, DoRA, and HypLoRA.

Dataset	Base Model	LoRA	DoRA	HypLoRA
AQuA	$0.17 \pm 0.03$	$0.17 \pm 0.03$	$0.19 \pm 0.02$	$0.11 \pm 0.01$
GSM8K	$0.16 \pm 0.03$	$0.20 \pm 0.03$	$0.19 \pm 0.03$	$0.11 \pm 0.02$
ARC-Challenge	$0.17 \pm 0.02$	$0.21 \pm 0.01$	$0.17 \pm 0.02$	$0.20 \pm 0.02$
Winogrande	$0.16 \pm 0.02$	$0.16 \pm 0.02$	$0.21 \pm 0.01$	$0.12 \pm 0.01$
OpenbookQA	$0.17 \pm 0.03$	$0.16 \pm 0.02$	$0.17 \pm 0.03$	$0.11 \pm 0.01$

This diverse collection allows us to investigate whether the hyperbolic characteristics observed in mathematical reasoning extend to broader domains of human knowledge and reasoning. The inclusion of both challenging (ARC-Challenge, HellaSwag) and more accessible (ARC-Easy, BoolQ) datasets enables us to examine how task difficulty influences the underlying geometric structure of token embeddings.

The results presented in Figure 4 demonstrate that the power-law distribution of token frequencies and the inverse relationship between frequency and embedding norms persist across all commonsense reasoning datasets, with power-law exponents ranging from  $\gamma = 1.76$  to  $\gamma = 1.94$ . Notably, the Social IQA dataset exhibits a slightly lower exponent ( $\gamma = 1.76$ ), suggesting that social reasoning tasks may have a somewhat different hierarchical structure, possibly due to the more nuanced and context-dependent nature of social interactions compared to factual or physical reasoning. Despite this variation, the overall pattern remains consistent: abstract concepts and function words maintain smaller norms and higher frequencies, while specific entities, proper nouns, and domain-specific terminology are positioned at greater distances from the origin.

### A.3 Hyperbolicity in the Final Hidden Layer of LLMs

In this part, we further present the analysis of the hyperbolicity of the hidden states in Tables 6 and Table 7. Considering five distinct reasoning datasets, including two mathematical reasoning datasets (AQuA and GSM8K) as well as three commonsense reasoning datasets (ARC-Challenge, Winogrande, and OpenbookQA), we observe that the base models consistently exhibit less hyperbolic structure (i.e., higher  $\delta$  values) in their final hidden layer representations compared to their initial token embeddings.

LoRA and DoRA generally reduce the  $\delta$  values, while the proposed HypLoRA method mostly achieves even lower values, indicating a higher degree of hyperbolicity in the learned representations. This effect is observed across most datasets in both model families. These empirical findings complement our analysis of the initial token embeddings: while the pretrained models begin with a latent hierarchical structure, as evidenced by hyperbolicity in the input layer, fine-tuning methods can either preserve or distort this property. The consistently lower  $\delta$  values of HypLoRA provide strong empirical evidence that our method actively preserves and enhances the hierarchical structure of the representations throughout the model, aligning the final contextualized embeddings with the geometric biases that are beneficial for reasoning.

## B Hyperbolicity on Different Metric Spaces

Table 2 presents the hyperbolicity values in both continuous (i.e., sphere space) and discrete metric spaces (i.e., tree, scale-free, and random graphs). We employ a consistent processing method similar to that used in Section 4 for embedding spaces. Specifically, we sample 1,000 four-tuples, compute the  $\delta$  value for each, and then take the maximum value.

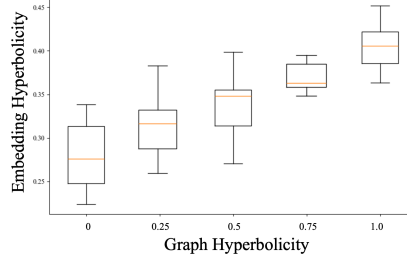


Figure 5: Empirical correlation between the ground-truth  $\delta$ -hyperbolicity of several reference graphs (tree, scale-free, PubMed, dense, sphere) and the  $\delta$  measured after embedding them with a two-layer GCN into Euclidean space; each point averages 1,000 sampled quadruples.

For the sphere space, we use a two-dimensional model and calculate hyperbolicity based on geodesic distances. The PubMed graph is sourced from Sen et al. [99]. The tree and dense graphs are generated using NetworkX [100]. For these graphs, we remove isolated nodes before performing our calculations to ensure consistency. We use the shortest-path distance on each graph as the distance measure, analogous to the concept of geodesics in continuous spaces.

In this study, we utilize the Euclidean distance to compute the hyperbolicity of token embeddings, following the approach proposed by [16]. To further assess the validity of this method, we embed graphs with varying degrees of hyperbolicity into Euclidean space using a two-layer GCN and compute hyperbolicity based on the distances between embeddings. The results, presented in Figure 5, indicate a positive correlation between the hyperbolicity of the original graphs and that of the embeddings, although the values do not exactly coincide. Building on this observed relationship, we calculate the hyperbolicity of token embeddings as a proxy for estimating their underlying geometric structure. In this context, lower hyperbolicity values suggest a more tree-like geometric configuration.

## C Exponential and Logarithmic Maps

The exponential and logarithmic maps are fundamental tools for navigating between the tangent space and the hyperbolic manifold. These maps enable us to perform computations in the familiar Euclidean tangent space while preserving the geometric properties of hyperbolic space.

### C.1 Exponential Map

The exponential map  $\exp_{\mathbf{x}}^K : \mathcal{T}_{\mathbf{x}}\mathcal{L}_K^n \rightarrow \mathcal{L}_K^n$  projects a tangent vector  $\mathbf{v} \in \mathcal{T}_{\mathbf{x}}\mathcal{L}_K^n$  at point  $\mathbf{x}$  onto the hyperboloid  $\mathcal{L}_K^n$ . Geometrically, it maps  $\mathbf{v}$  to the point  $\exp_{\mathbf{x}}^K(\mathbf{v}) := \gamma(1)$ , where  $\gamma$  is the unique geodesic satisfying  $\gamma(0) = \mathbf{x}$  and  $\dot{\gamma}(0) = \mathbf{v}$ .

The exponential map is given by:

$$\exp_{\mathbf{x}}^K(\mathbf{v}) = \cosh\left(\frac{\|\mathbf{v}\|_{\mathcal{L}}}{\sqrt{K}}\right) \mathbf{x} + \sqrt{K} \sinh\left(\frac{\|\mathbf{v}\|_{\mathcal{L}}}{\sqrt{K}}\right) \frac{\mathbf{v}}{\|\mathbf{v}\|_{\mathcal{L}}}, \quad (11)$$

where  $\|\mathbf{v}\|_{\mathcal{L}} = \sqrt{\langle \mathbf{v}, \mathbf{v} \rangle_{\mathcal{L}}}$  is the norm of the tangent vector under the Lorentzian inner product.

At the origin  $\mathbf{o} = (\sqrt{K}, 0, \dots, 0)$ , for a tangent vector  $\mathbf{v} = (0, \mathbf{u})$  where  $\mathbf{u} \in \mathbb{R}^n$ , the exponential map simplifies to:

$$\exp_{\mathbf{o}}^K(\mathbf{v}) = \left( \sqrt{K} \cosh\left(\frac{\|\mathbf{u}\|}{\sqrt{K}}\right), \sqrt{K} \sinh\left(\frac{\|\mathbf{u}\|}{\sqrt{K}}\right) \frac{\mathbf{u}}{\|\mathbf{u}\|} \right). \quad (12)$$

## C.2 Logarithmic Map

The logarithmic map  $\log_{\mathbf{x}}^K : \mathcal{L}_K^n \rightarrow \mathcal{T}_{\mathbf{x}}\mathcal{L}_K^n$  is the inverse of the exponential map. It projects a point  $\mathbf{y} \in \mathcal{L}_K^n$  back to the tangent space at  $\mathbf{x}$ :

$$\log_{\mathbf{x}}^K(\mathbf{y}) = \frac{\cosh^{-1}\left(-\frac{\langle \mathbf{x}, \mathbf{y} \rangle_{\mathcal{L}}}{K}\right)}{\sqrt{\left(\frac{\langle \mathbf{x}, \mathbf{y} \rangle_{\mathcal{L}}}{K}\right)^2 - 1}} \left( \mathbf{y} + \frac{\langle \mathbf{x}, \mathbf{y} \rangle_{\mathcal{L}}}{K} \mathbf{x} \right). \quad (13)$$

These maps satisfy the inverse relationships:  $\log_{\mathbf{x}}^K(\exp_{\mathbf{x}}^K(\mathbf{v})) = \mathbf{v}$  and  $\exp_{\mathbf{x}}^K(\log_{\mathbf{x}}^K(\mathbf{y})) = \mathbf{y}$ .

## C.3 Notation in the Main Text

In the main text, we use the shorthand notation  $\Pi_{\exp}^K$  and  $\Pi_{\log}^K$  to denote general projection operators between Euclidean space  $\mathbb{R}^n$  and hyperbolic space  $\mathcal{L}_K^n$ . The exponential and logarithmic maps described above represent one valid instantiation of these operators:

$$\Pi_{\exp}^K(\mathbf{x}) := \exp_{\mathbf{o}}^K((0, \mathbf{x})), \quad (14)$$

$$\Pi_{\log}^K(\mathbf{y}) := \log_{\mathbf{o}}^K(\mathbf{y})_{[1:]}, \quad (15)$$

where  $(0, \mathbf{x}) \in \mathbb{R}^{1+d}$  denotes the vector obtained by prepending a zero to  $\mathbf{x} \in \mathbb{R}^d$ , and  $(\cdot)_{[1:]}$  denotes the restriction to the last  $d$  coordinates (i.e., removal of the first coordinate). However, other diffeomorphisms [101] between Euclidean and hyperbolic spaces. The choice of projection method can be adapted based on computational efficiency and numerical stability requirements, while the core principle of our approach, performing the low-rank transformation directly on the hyperbolic manifold, remains unchanged.

**Important Observation.** Regardless of the specific projection method used, when these maps are applied consecutively at the same base point without intermediate operations on the manifold, they effectively cancel each other out. For example,  $\log_{\mathbf{o}}^K(\exp_{\mathbf{o}}^K(\mathbf{v})) = \mathbf{v}$ . This is why the conventional tangent-space approach for hyperbolic neural networks [33, 32] does not directly benefit LLM adaptation, where the hyperbolic geometry is effectively bypassed. Our Direct Lorentz Low-Rank Transformation (LLR) addresses this limitation by operating directly on the hyperbolic manifold between the projection steps, ensuring that the geometric properties of hyperbolic space are preserved and utilized.

## D Lorentz Transformation

In the context of special relativity, Lorentz transformations are linear mappings that preserve the spacetime interval between events, ensuring the constancy of the speed of light across all inertial frames. These transformations can be categorized into two primary types: Lorentz boosts and Lorentz rotations [102, 103].

### D.1 Lorentz Boost

A Lorentz boost corresponds to a transformation between two inertial reference frames moving at a constant relative velocity. Given a velocity vector  $\mathbf{v} \in \mathbb{R}^n$  with magnitude  $\|\mathbf{v}\| < 1$ , the Lorentz boost matrix  $\mathbf{B}$  mixes time and space coordinates:

$$\mathbf{B} = \begin{bmatrix} \gamma & -\gamma \mathbf{v}^{\top} \\ -\gamma \mathbf{v} & \mathbf{I} + \frac{\gamma^2}{1+\gamma} \mathbf{v} \mathbf{v}^{\top} \end{bmatrix}, \quad (16)$$

where  $\gamma = \frac{1}{\sqrt{1-\|\mathbf{v}\|^2}}$  is the Lorentz factor.

### D.2 Lorentz Rotation

A Lorentz rotation involves only the rotation of spatial coordinates while preserving the time coordinate:

$$\mathbf{R} = \begin{bmatrix} 1 & \mathbf{0}^{\top} \\ \mathbf{0} & \tilde{\mathbf{R}} \end{bmatrix}, \quad (17)$$

where  $\tilde{\mathbf{R}} \in SO(n)$  is a spatial rotation matrix.

**Our Spatial-like Transformation.** In our Direct Lorentz Low-Rank Transformation (LLR), we apply transformations exclusively to the spatial components while maintaining the constraint of the Lorentz manifold. Given a point  $\mathbf{x}^H = (x_0^H, \mathbf{x}_s^H) \in \mathcal{L}_K^n$ , our transformation is:

$$\mathbf{LLR}(BA, \mathbf{x}^H) = (\sqrt{\|BA\mathbf{x}_s^H\|^2 + K}, BA\mathbf{x}_s^H), \quad (18)$$

where we transform the spatial component  $\mathbf{x}_s^H$  and recompute the time component to maintain the Lorentz constraint  $x_0^2 - \|\mathbf{x}_s\|^2 = K$ .

This can be decomposed into two sequential transformations:

$$\mathbf{y}^H = (y_0^H, \mathbf{y}_s^H) = (\sqrt{\|A\mathbf{x}_s^H\|^2 + K}, A\mathbf{x}_s^H), \quad (19)$$

$$\mathbf{z}^H = (z_0^H, \mathbf{z}_s^H) = (\sqrt{\|B\mathbf{y}_s^H\|^2 + K}, B\mathbf{y}_s^H). \quad (20)$$

**Interpretation as a Constrained Lorentz Rotation.** Our transformation can be viewed as a special case of Lorentz rotation where: (1) We apply a linear transformation to the spatial coordinates:  $\mathbf{x}_s^H \mapsto BA\mathbf{x}_s^H$ ; (2) We recompute the time component to preserve the manifold constraint:  $x_0^H \mapsto \sqrt{\|BA\mathbf{x}_s^H\|^2 + K}$ . This approach differs from a standard Lorentz rotation in two ways (see also [37]): (1) the spatial transformation  $BA$  is not necessarily orthogonal (i.e.,  $BA \notin SO(n)$ ); (2) the time component is not preserved but rather recomputed to maintain the manifold constraint.

In matrix form, our transformation can be expressed as:

$$\begin{bmatrix} z_0^H \\ \mathbf{z}_s^H \end{bmatrix} = \begin{bmatrix} \frac{\sqrt{\|BA\mathbf{x}_s^H\|^2 + K}}{\sqrt{\|\mathbf{x}_s^H\|^2 + K}} & \mathbf{0}^\top \\ \mathbf{0} & BA \end{bmatrix} \begin{bmatrix} x_0^H \\ \mathbf{x}_s^H \end{bmatrix} \quad (21)$$

The key property is that this transformation preserves the Lorentz manifold structure: if  $\mathbf{x}^H \in \mathcal{L}_K^n$ , then  $\mathbf{LLR}(BA, \mathbf{x}^H) \in \mathcal{L}_K^n$ , as verified by:

$$(z_0^H)^2 - \|\mathbf{z}_s^H\|^2 = \|BA\mathbf{x}_s^H\|^2 + K - \|BA\mathbf{x}_s^H\|^2 = K. \quad (22)$$

This spatial-like transformation approach allows us to leverage the low-rank structure of  $BA$  while maintaining the geometric properties of the hyperbolic space, providing a computationally efficient method for hyperbolic low-rank adaptation.

## E Transformation Analysis

This section provides a detailed analysis of how HypLoRA differs from standard LoRA by examining the higher-order terms introduced through hyperbolic geometry.

*Proof.* Let  $\mathbf{x} \in \mathbb{R}^d$  be an input token embedding. Let  $A \in \mathbb{R}^{r \times d}$  and  $B \in \mathbb{R}^{k \times r}$  be low-rank matrices with rank  $r \ll \min\{d, k\}$ . Consider the  $d$ -dimensional hyperbolic space  $\mathcal{L}_K^d$  (Lorentz model) with curvature  $C = -1/K$ , where  $K > 0$ .

Our goal is to analyze how the HypLoRA update differs from the LoRA update and to understand the impact of token norms  $\|\mathbf{x}\|$  on the higher-order terms introduced by HypLoRA.

**Mapping the Input Embedding to Hyperbolic Space.** Following previous work [32], we interpret the Euclidean token embedding  $\mathbf{x}$  as an element in the tangent space at the origin  $\mathbf{o}$  of the hyperbolic space  $\mathcal{L}_K^d$ . The tangent vector is given by  $\mathbf{v} = (0, \mathbf{x}) \in T_{\mathbf{o}}\mathcal{L}_K^d$ . The exponential map  $\exp_{\mathbf{o}}^K$  projects  $\mathbf{v}$  onto the hyperbolic space:

$$\exp_{\mathbf{o}}^K(\mathbf{v}) = \left( \sqrt{K} \cosh\left(\frac{\|\mathbf{v}\|_{\mathcal{L}}}{\sqrt{K}}\right), \sqrt{K} \sinh\left(\frac{\|\mathbf{v}\|_{\mathcal{L}}}{\sqrt{K}}\right) \frac{\mathbf{v}}{\|\mathbf{v}\|_{\mathcal{L}}} \right), \quad (23)$$

where  $\|\mathbf{v}\|_{\mathcal{L}}$  denotes the Minkowski norm. Since  $\mathbf{v} = (0, \mathbf{x})$  and  $\|\mathbf{v}\|_{\mathcal{L}} = \|\mathbf{x}\|$ , the exponential map simplifies to:

$$\exp_{\mathbf{o}}^K(\mathbf{v}) = \left( \sqrt{K} \cosh\left(\frac{\|\mathbf{x}\|}{\sqrt{K}}\right), \sqrt{K} \sinh\left(\frac{\|\mathbf{x}\|}{\sqrt{K}}\right) \frac{\mathbf{x}}{\|\mathbf{x}\|} \right). \quad (24)$$

**Approximations.** For small  $\frac{\|\mathbf{x}\|}{\sqrt{K}}$ , let  $z = \|\mathbf{x}\|$  we can use the Taylor series expansions:

$$\cosh\left(\frac{z}{\sqrt{K}}\right) \approx 1 + \frac{z^2}{2K}, \quad \sinh\left(\frac{z}{\sqrt{K}}\right) \approx \frac{z}{\sqrt{K}} + \frac{z^3}{6K^{3/2}}. \quad (25)$$

Applying these to the exponential map of  $\mathbf{x}$ :

$$u_0^H \approx \sqrt{K} + \frac{\|\mathbf{x}\|^2}{2\sqrt{K}}, \quad (26)$$

$$\mathbf{u}_{\text{space}}^H \approx \mathbf{x} + \frac{\|\mathbf{x}\|^2}{6K} \mathbf{x}. \quad (27)$$

**Applying Low-Rank Transformations to the Approximated Embedding.** Using the approximated  $\mathbf{u}_{\text{space}}^H$ , we apply the transformations.

First transformation:

$$\mathbf{y}_{\text{space}}^H = A\mathbf{u}_{\text{space}}^H \approx A\left(\mathbf{x} + \frac{\|\mathbf{x}\|^2}{6K} \mathbf{x}\right) = A\mathbf{x} + \frac{\|\mathbf{x}\|^2}{6K} A\mathbf{x}. \quad (28)$$

Second transformation:

$$\mathbf{z}_{\text{space}}^H = B\mathbf{y}_{\text{space}}^H \approx BA\mathbf{x} + \frac{\|\mathbf{x}\|^2}{6K} BA\mathbf{x}. \quad (29)$$

Compute the time component after the transformations:

$$z_0^H = \sqrt{K + \|\mathbf{z}_{\text{space}}^H\|^2}. \quad (30)$$

**Approximating the Logarithmic Map.** We map the transformed hyperbolic point  $\mathbf{z}^H = (z_0^H, \mathbf{z}_{\text{space}}^H)$  back to the tangent space at the origin using the logarithmic map  $\log_o^K$ :

$$\Delta Q^{\text{Hyp}} = \log_o^K(\mathbf{z}^H) = \sqrt{K} \cdot \text{arcosh}\left(\frac{z_0^H}{\sqrt{K}}\right) \frac{\mathbf{z}_{\text{space}}^H}{\sqrt{(z_0^H)^2 - K}}. \quad (31)$$

Using the approximation  $z_0^H \approx \sqrt{K} + \frac{\|\mathbf{z}_{\text{space}}^H\|^2}{2\sqrt{K}}$  and for small  $\delta = \frac{\|\mathbf{z}_{\text{space}}^H\|^2}{2K}$ , we have:

$$\text{arcosh}\left(\frac{z_0^H}{\sqrt{K}}\right) \approx \text{arcosh}(1 + \delta) \approx \sqrt{2\delta} = \frac{\|\mathbf{z}_{\text{space}}^H\|}{\sqrt{K}}, \quad (32)$$

$$\sqrt{(z_0^H)^2 - K} \approx \|\mathbf{z}_{\text{space}}^H\|. \quad (33)$$

Therefore, the logarithmic map simplifies to:

$$\Delta Q^{\text{Hyp}} \approx \mathbf{z}_{\text{space}}^H. \quad (34)$$

**Comparing HypLoRA and LoRA Updates.** The HypLoRA update is:

$$\Delta Q^{\text{Hyp}} \approx BA\mathbf{x} + \frac{\|\mathbf{x}\|^2}{6K} BA\mathbf{x}. \quad (35)$$

The LoRA update is:

$$\Delta Q^{\text{LoRA}} = BA\mathbf{x}. \quad (36)$$

The difference between the updates is:

$$\Delta Q^{\text{Hyp}} - \Delta Q^{\text{LoRA}} = \frac{\|\mathbf{x}\|^2}{6K} BA\mathbf{x}. \quad (37)$$



**Impact of Token Norms on Higher-Order Terms.** The higher-order term  $\frac{\|\mathbf{x}\|^2}{6K} B A \mathbf{x}$  is proportional to  $\|\mathbf{x}\|^2$ . Since  $\|\mathbf{x}\|$  reflects the specificity of the token in the hierarchical structure (larger norms correspond to more specific tokens), this term becomes significant for tokens representing specific concepts.

**Impact on Attention Scores.** The HypLoRA attention scores are computed as:

$$\text{Scores}_{\text{HypLoRA}} = \frac{(Q^{\text{orig}} + \Delta Q^{\text{Hyp}})(K^{\text{orig}} + \Delta K^{\text{Hyp}})^{\top}}{\sqrt{d_k}}, \quad (38)$$

where  $\Delta K^{\text{Hyp}}$  is derived similarly.

The difference in attention scores includes higher-order terms dependent on  $\|\mathbf{x}\|^2$ :

$$\Delta \text{Scores} = \text{Scores}_{\text{HypLoRA}} - \text{Scores}_{\text{LoRA}}. \quad (39)$$

These higher-order terms enable HypLoRA to capture more complex hierarchical relationships, particularly for tokens with larger norms. □

**Remark 1. Alignment with Token Hierarchy:** The higher-order terms in HypLoRA’s updates are proportional to  $\|\mathbf{x}\|^2$ , correlating with the specificity of tokens in the hierarchical structure. As a result, HypLoRA places greater emphasis on more specific tokens, enhancing its ability to model detailed relationships.

**Role of Curvature  $C$ :** The curvature  $C = -1/K$  scales the higher-order corrections. Smaller  $K$  (larger negative curvature) amplifies these terms, aligning with the hyperbolic nature of token embeddings. In practice, the curvature parameter  $K$  can be tuned to ensure this condition is satisfied for typical token embedding norms.

**Effectiveness of HypLoRA:** By incorporating these higher-order terms, HypLoRA leverages the inherent hierarchical and hyperbolic structure of token embeddings. This leads to improved performance, especially on problems requiring complex reasoning, explaining why the proposed method performs better on more challenging datasets.

## F Full Comparison

While the main body of our paper focuses on comparing HypLoRA against the standard LoRA baseline to demonstrate the core effectiveness of our hyperbolic fine-tuning approach, this section provides a comprehensive evaluation against a broader range of parameter-efficient fine-tuning methods, such as Prefix tuning [68], Series and Parallel adapters [70], and DoRA [75], providing a more complete picture of HypLoRA’s performance relative to the current landscape of efficient fine-tuning techniques. This extended comparison validates that our improvements are not merely due to increased model capacity or specific architectural choices, but rather stem from the fundamental advantages of incorporating hyperbolic geometry into the adaptation process.

### F.1 Implementation Details

To ensure consistency and comparability, our experimental setup closely followed the training configurations outlined in Hu et al. [98]. Across all fine-tuning tasks, we employed the AdamW optimizer with a learning rate of  $3 \times 10^{-4}$  and trained for a total of three epochs. LoRA modules (and consequently, HypLoRA adapters) were integrated into both the Multi-Head Attention (MHA) and MLP layers of the foundation models. A key hyperparameter for HypLoRA is the curvature  $K$  (defining the hyperbolic curvature as  $-1/K$ ), which was initialized by searching the set  $\{0.5, 1.0\}$ . For evaluation, final scores were micro-averaged for arithmetic reasoning and averaged for commonsense reasoning across the datasets, thereby giving equal weight to each individual prompt, regardless of the varying number of questions per dataset (e.g., 1, 319 in GSM8K versus 238 in MAWPS).

For baseline methods, we adopted the following approach: results for Prefix tuning [68], Series adapters, and Parallel adapters [70] are directly cited from Hu et al. [98] to ensure fair comparison

Table 8: Comprehensive comparison of parameter-efficient fine-tuning methods on mathematical reasoning tasks. Results marked with \* are from [98], while † indicates our reproduced results. The percentage following each dataset name indicates the proportion of prompts relative to the total number of inference prompts. M.AVG represents the micro-average accuracy across all datasets. Best results for each model are highlighted in bold. OOT indicates out-of-time during training.

Base Model	PEFT Method	MAWPS(8.5%)	SVAMP(35.6%)	GSM8K(46.9%)	AQuA(9.0%)	M.AVG
GPT-3.5	None	87.4	69.9	56.4	38.9	62.3
LLaMA-7B	None	51.7	32.4	15.7	16.9	24.8
	Prefix*	63.4	38.1	24.4	14.2	31.7
	Series*	77.7	52.3	33.3	15.0	42.2
	Parallel*	82.4	49.6	35.3	18.1	42.8
	LoRA*	79.0	52.1	37.5	18.9	44.6
	LoRA†	81.9	48.2	38.3	18.5	43.7
	DoRA	80.0	48.8	39.0	16.4	43.9
	<b>HypLoRA (Ours)</b>	<b>79.0</b>	<b>49.1</b>	<b>39.1</b>	<b>20.5</b>	<b>44.4</b>
LLaMA-13B	None	65.5	37.5	32.4	15.0	35.5
	Prefix*	66.8	41.4	31.1	15.7	36.4
	Series*	78.6	50.8	44.0	22.0	47.4
	Parallel*	81.1	55.7	43.3	20.5	48.9
	LoRA*	83.6	54.6	47.5	18.5	50.5
	LoRA†	83.5	54.7	48.5	18.5	51.0
	DoRA	83.0	54.6	OOT	18.9	NA
	<b>HypLoRA (Ours)</b>	<b>83.2</b>	<b>54.8</b>	<b>49.0</b>	<b>21.5</b>	<b>51.5</b>
Gemma-7B	None	76.5	60.4	38.4	25.2	48.3
	LoRA	91.6	76.2	66.3	28.9	68.6
	DoRA	90.7	79.2	68.3	33.9	71.0
	<b>HypLoRA (Ours)</b>	<b>89.5</b>	<b>78.7</b>	<b>69.5</b>	<b>32.7</b>	<b>71.2</b>
LLaMA3-8B	None	79.8	50.0	54.7	21.0	52.1
	LoRA	92.7	78.9	70.8	30.4	71.9
	DoRA	90.3	79.8	73.3	21.3	72.4
	<b>HypLoRA (Ours)</b>	<b>91.6</b>	<b>80.5</b>	<b>74.0</b>	<b>34.2</b>	<b>74.2</b>
Gemma3-4B	LoRA	90.8	77.3	72.3	50.8	73.7
	DoRA	89.5	78.8	68.5	52.4	72.5
	<b>HypLoRA (Ours)</b>	<b>88.2</b>	<b>83.9</b>	<b>76.1</b>	<b>53.2</b>	<b>77.8</b>
Qwen2.5-7B	LoRA	90.8	84.4	78.6	68.1	80.8
	DoRA	92.8	87.4	80.4	64.2	82.5
	<b>HypLoRA (Ours)</b>	<b>91.2</b>	<b>92.2</b>	<b>87.9</b>	<b>71.6</b>	<b>88.3</b>

under identical experimental conditions. For LoRA and DoRA, we conducted independent reimplementations following their respective original papers and parameters [31, 75] to enable rigorous and controlled comparisons.

## F.2 Comparison on Mathematical Reasoning

Looking at the mathematical reasoning comparison table, several key experimental findings emerge regarding HypLoRA’s performance across different model architectures and datasets. The results demonstrate that HypLoRA consistently outperforms standard LoRA across multiple model families, with particularly notable improvements on more challenging datasets. For the Gemma-7B model, HypLoRA achieves a micro-averaged accuracy of 71.2%, surpassing LoRA’s 68.6%. For LLaMA3-8B, HypLoRA reaches 74.2% compared to LoRA’s 71.9%. The improvements are especially pronounced on the AQuA dataset, which requires complex algebraic reasoning. Specifically, HypLoRA shows gains of 3.8 percentage points over LoRA on Gemma-7B (32.7% vs 28.9%) and 3.8 points on LLaMA3-8B (34.2% vs 30.4%). This pattern suggests that HypLoRA’s hyperbolic geometry is particularly effective for problems requiring multi-step reasoning and understanding of hierarchical mathematical relationships.

The consistency of improvements across different model architectures further validates the generalizability of the hyperbolic approach. While HypLoRA shows competitive performance on simpler datasets like MAWPS, the performance advantages become more significant on challenging datasets like GSM8K and AQuA, which demand sophisticated reasoning capabilities. For instance, on GSM8K, HypLoRA achieves 69.5% accuracy on Gemma-7B versus 66.3% for LoRA, and 74.0%

Table 9: Extended commonsense reasoning accuracy (%) for GPT-3.5 and for LoRA, DoRA, and HypLoRA on LLaMA3-8B, Gemma3-4B, and Qwen2.5-7B. Columns correspond to BoolQ, PIQA, SIQA, HellaSwag, WinoGrande, ARC-e, ARC-c, and OBQA; the rightmost column reports the macro average across the eight benchmarks.

Base Model	PEFT Method	# Params (%)	BoolQ	PIQA	SIQA	HellaSwag	WinoGrande	ARC-e	ARC-c	OBQA	AVG
GPT-3.5	None	None	73.1	85.4	68.5	78.5	66.1	89.8	79.9	74.8	77.0
LLaMA3-8B	LoRA	0.70	70.8	85.2	79.9	91.7	84.3	84.2	71.2	79.0	80.8
	DoRA	0.71	72.1	85.5	79.6	92.8	83.3	85.2	72.1	84.0	81.8
	<b>HypLoRA (Ours)</b>	0.70	<b>74.1</b>	<b>87.6</b>	<b>80.6</b>	<b>94.5</b>	<b>84.7</b>	<b>90.4</b>	<b>81.2</b>	<b>85.2</b>	<b>84.8</b>
Gemma3-4B	LoRA	1.04	68.1	83.2	77.2	88.9	<b>80.5</b>	84.5	69.9	83.6	79.5
	DoRA	1.05	68.1	84.3	78.4	88.3	80.1	84.1	70.8	83.8	79.7
	<b>HypLoRA (Ours)</b>	1.04	<b>70.0</b>	<b>84.3</b>	<b>79.2</b>	<b>91.5</b>	80.3	<b>89.1</b>	<b>75.9</b>	<b>86.4</b>	<b>82.5</b>
Qwen2.5-7B	LoRA	0.71	<b>73.4</b>	<b>89.5</b>	79.5	93.6	84.1	92.8	82.0	87.0	85.2
	DoRA	0.72	71.7	88.7	79.0	93.7	84.1	92.4	82.8	88.4	85.1
	<b>HypLoRA (Ours)</b>	0.71	72.8	89.3	<b>79.8</b>	<b>94.8</b>	<b>84.4</b>	<b>95.5</b>	<b>87.5</b>	<b>90.8</b>	<b>87.0</b>

on LLaMA3-8B versus 70.8% for LoRA. These correspond to gains of 3.2 points over LoRA on both Gemma-7B and LLaMA3-8B. Notably, on the most recent models, HypLoRA demonstrates substantial improvements: on Gemma3-4B, HypLoRA achieves 77.8% M.AVG compared to LoRA’s 73.7% (+4.1 points), and on Qwen2.5-7B, HypLoRA reaches 88.3% versus LoRA’s 80.8% (+7.5 points). The fact that HypLoRA maintains superior performance across both older (LLaMA-7B, LLaMA-13B) and newer (LLaMA3-8B, Gemma3-4B, Qwen2.5-7B) model architectures demonstrates the robustness of incorporating hyperbolic inductive biases into parameter-efficient fine-tuning, regardless of the underlying model’s specific architectural details or training paradigms.

### F.3 Comparison on Commonsense Reasoning

HypLoRA demonstrates substantial improvements over standard LoRA across diverse commonsense reasoning benchmarks, as shown in Table 9. The commonsense reasoning tasks evaluated include BoolQ (yes/no question answering), PIQA (physical commonsense inference), SIQA (social interaction reasoning), HellaSwag (commonsense natural language inference), WinoGrande (pronoun disambiguation), ARC-e and ARC-c (science question answering with easy and challenging difficulty), and OBQA (open book question answering). These benchmarks collectively assess the model’s ability to understand implicit knowledge, contextual nuances, and real-world reasoning patterns. The consistent performance gains across all three model architectures and eight diverse benchmarks indicate that the hierarchical inductive bias introduced by hyperbolic geometry effectively captures the implicit relational structures underlying commonsense reasoning.

### F.4 GPU Usage

Table 10 presents a comprehensive comparison of memory usage across different fine-tuning methods for both LLaMA3-8B and Gemma3-4B models. The results demonstrate that HypLoRA maintains comparable memory efficiency to the baseline LoRA method. Specifically, HypLoRA with stereographic projection achieves identical memory consumption to LoRA (30.12 GB for LLaMA3-8B and 14.61 GB for Gemma3-4B), while HypLoRA with exponential/logarithmic maps introduces only a minimal overhead of 0.02 GB for LLaMA3-8B and 0.01 GB for Gemma3-4B. Notably, both HypLoRA variants significantly outperform DoRA in terms of memory efficiency, with DoRA requiring 30.23 GB and 14.62 GB, respectively. These results confirm that our hyperbolic adaptation approach does not compromise memory efficiency while delivering superior performance improvements, making HypLoRA a practical choice for resource-constrained environments where both performance gains and memory conservation are critical considerations.

Table 10: Allocated Memory Usage Comparison for Fine-tuning Methods

Method	LLaMA3-8B		Gemma3-4B	
	Param (%)	Memory (GB)	Param (%)	Memory (GB)
LoRA	0.7002	30.12	1.0371	14.61
DoRA	0.7098	30.23	1.0516	14.62
HypLoRA	0.7099	30.14	1.0518	14.62

Table 11: Example problem with solutions from LoRA and HypLoRA (Case 1)

<b>Question</b>	To have the car for the weekend, Wilson’s report card needs to show that he received 80 or higher in his math class. His math grade is calculated by the average of 5 test scores. On those tests, he received: 65, 94, 81, 86, and 74. What is his current math grade? <b>Answer:</b> 80.0
<b>LoRA</b>	To find Wilson’s current math grade, we need to take the average of his 5 test scores. First, we add up all 5 test scores: $65 + 94 + 81 + 86 + 74 = 380$ . Then, we divide the sum by the number of tests: $380 \div 5 = 76$ . Therefore, Wilson’s current math grade is 76. <b>Answer:</b> 76.
<b>HypLoRA</b>	To find Wilson’s current math grade, we need to take the average of his 5 test scores. To do this, we add up the scores and divide by 5: $65 + 94 + 81 + 86 + 74 = 400$ . Then, we divide by 5: $400 \div 5 = 80$ . Therefore, Wilson’s current math grade is 80. <b>Answer:</b> 80.

## G Case Study

In this section, we present three examples to demonstrate the advantages of incorporating HypLoRA over standard LoRA fine-tuning.

Table 12: Example problem with solutions from LoRA and HypLoRA (Case 2)

<b>Question</b>	It takes 240 minutes of walking to break in a new pair of shoes. Jason wants to try out for the track team in three weeks. If he can walk 4 days a week to break in the new shoes, how long does he have to spend walking each day? <b>Answer:</b> 20.0 minutes
<b>LoRA</b>	To find out how long Jason has to spend walking each day, we need to divide the total number of minutes it takes to break in the shoes by the number of days he has to do it. Jason has 240 minutes divided by 4 days per week. This results in 60 minutes per day. <b>Answer:</b> 60.
<b>HypLoRA</b>	First, we need to find how many minutes Jason needs to walk in total to break in the new shoes: 240 minutes. Next, we find how many minutes Jason needs to walk each week over three weeks: 240 minutes divided by 3 weeks equals 80 minutes per week. Then, to find how many minutes he needs to walk each day: 80 minutes per week divided by 4 days equals 20 minutes per day. <b>Answer:</b> 20.

Table 13: Example problem with solutions from LoRA and HypLoRA (Case 3)

<b>Question</b>	Gissela, Gordy, and Gary are truck drivers. Gissela has a truck large enough to haul 4,000 pounds of gravel. Gordy’s truck can haul 800 pounds more than Gissela’s truck. When Gary brings his truck and joins Gissela and Gordy, the three trucks combined can haul a total of 11,600 pounds of gravel. How many pounds of gravel can Gary’s truck carry? <b>Answer:</b> 2,800 pounds
<b>LoRA</b>	Let’s start by assigning variables to represent the amount of gravel each truck can haul: Gissela’s truck hauls 4,000 pounds. Gordy’s truck hauls $4,000 + 800 = 4,800$ pounds. Now, we set up an equation for the total amount of gravel the three trucks can haul: $4,000 + 4,800 + x = 11,600$ , where $x$ is the amount Gary’s truck hauls. Simplifying the equation: $9,800 + x = 11,600$ , we subtract 9,800 from both sides to get $x = 11,600 - 9,800 = 1,800$ . Therefore, Gary’s truck hauls 1,800 pounds. <b>Answer:</b> 1,800.
<b>HypLoRA</b>	Gissela’s truck can haul 4,000 pounds of gravel. Gordy’s truck can haul 800 more pounds than Gissela’s truck, so it can haul 4,800 pounds of gravel. Together, Gissela and Gordy’s trucks can haul 8,800 pounds of gravel. If the three trucks combined can haul 11,600 pounds, then Gary’s truck can haul $11,600 - 8,800 = 2,800$ pounds of gravel. <b>Answer:</b> 2,800.

These examples demonstrate how HypLoRA consistently provides more accurate reasoning compared to LoRA across different types of mathematical problems. In Case 1, LoRA drops 20 points when summing the five scores (reporting 380 instead of 400) and therefore produces the wrong average. This seemingly small arithmetic lapse aligns with the observation that LLMs often rely on high-level pattern similarity rather than exact computation [104]. By preserving greater separation among numerically close but semantically distinct tokens (e.g., 380 vs. 400), the hyperbolic representation in HypLoRA keeps the sequence of operations faithful and recovers the correct average.

In Case 2, LoRA immediately divides 240 minutes by the four weekly walking days, yielding 60 minutes per day and ignoring that the 240-minute budget must be spread over three weeks. HypLoRA

correctly reasons in stages: divide 240 by 3 weeks, then by 4 days per week, recovering the required 20 minutes per day and showing stronger temporal reasoning.

In Case 3, LoRA actually sets up the correct balance equation  $4,000 + 4,800 + x = 11,600$  but subtracts 9,800 from 11,600 rather than 8,800, reporting  $x = 1,800$ . HypLoRA carries the subtraction through correctly and outputs the true 2,800 pounds. Together, these examples illustrate how the hyperbolic geometry employed by HypLoRA enables better handling of multi-step reasoning, maintaining both semantic context and numerical consistency in mathematical problem-solving scenarios.

Overall, these cases highlight a consistent trend: LoRA frequently derails on either a single arithmetic step (Cases 1 and 3) or a latent multi-hop dependency (Case 2), whereas HypLoRA preserves each intermediate calculation, keeps quantities well separated in representation space, and consequently delivers the correct final answers. These qualitative observations complement the quantitative gains reported in the main paper.