

HoPE: Hyperbolic Rotary Positional Encoding for Stable Long-Range Dependency Modeling in Large Language Models

Chang Dai¹ Hongyu Shan² Mingyang Song³ Di liang⁴

¹Peking University ²Tianjin University

³Tencent ⁴Fudan University

daichang@pku.edu.cn, shhy@tju.edu.cn,

nickmysong@tencent.com, dliang@fudan.edu.cn

Abstract

Positional encoding mechanisms enable Transformers to model sequential structure and long-range dependencies in text. While absolute positional encodings struggle with extrapolation to longer sequences due to fixed positional representations, and relative approaches like Alibi exhibit performance degradation on extremely long contexts, the widely-used Rotary Positional Encoding (RoPE) introduces oscillatory attention patterns that hinder stable long-distance dependency modelling. We address these limitations through a geometric reformulation of positional encoding. Drawing inspiration from Lorentz transformations in hyperbolic geometry, we propose Hyperbolic Rotary Positional Encoding (**HoPE**), which leverages hyperbolic functions to implement Lorentz rotations on token representations. Theoretical analysis demonstrates that RoPE is a special case of our generalized formulation. HoPE fundamentally resolves RoPE’s slation issues by enforcing monotonic decay of attention weights with increasing token distances. Extensive experimental results, including perplexity evaluations under several extended sequence benchmarks, show that HoPE consistently exceeds existing positional encoding methods. These findings underscore HoPE’s enhanced capacity for representing and generalizing long-range dependencies. Data and code will be available.

1 Introduction

Positional encoding mechanisms strive to provide Transformers (Vaswani et al., 2023) with stable and expressive representations of the sequential structure, thereby addressing the order-agnostic nature of the multi-head attention module (Raffel et al., 2023; Anil et al., 2022). By encoding information about the relative (Shaw et al., 2018) or absolute positions of tokens, positional encodings enable models to capture the intricacies of syntactic and semantic dependencies across different spans of

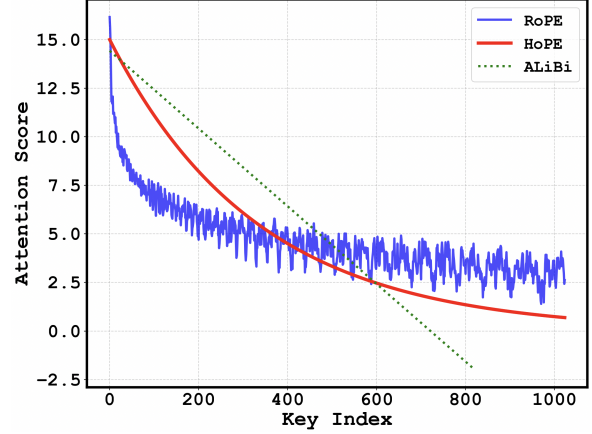


Figure 1: Illustration of attention scores. For the same embedding, when using the RoPE, alibi, and HoPE methods to represent positional information, this shows the corresponding changing trends of the Attention scores as the distance varies.

text (Wang and Chen, 2020). Without such positional signals, Transformers can struggle to fully delineate word-order information and effectively leverage long-range context (Haviv et al., 2022). A variety of strategies (Ruoss et al., 2023; Kazemnejad et al., 2023; Li et al., 2024; Chen et al., 2023; Xiong et al., 2023) have been proposed to incorporate positional knowledge in large language models (LLMs), aiming to ensure reliable and generalizable sequence representations (Chowdhury and Caragea, 2023; Sun et al., 2022).

Absolute positional encodings (Devlin et al., 2019), which typically employ sinusoidal signals or learnable embeddings indexed by token positions, are straightforward to implement but often struggle with length extrapolation, as their fixed position representations do not naturally extend to unseen sequence lengths. In contrast, some approaches (Press et al., 2022; Chi et al., 2022a, 2023), which introduce a distance-based attention bias, demonstrate improved performance over absolute encodings but can still degrade when sequences become

very long, revealing limitations in capturing stable correlations at distant positions. More recently, Rotary Positional Encoding (RoPE) (Su et al., 2023) has gained substantial traction by rotating query and key vectors at various frequencies as a function of token positions (Barbero et al., 2024). However, RoPE exhibits an oscillatory attention pattern, in which attention weights fluctuate rather than decrease smoothly as the token distance increases, making it difficult to reliably represent long-distance dependencies, as shown in figure 1.

Recent applications demonstrate that RoPE is widely adopted in many state-of-the-art large-scale models, such as Llama, Gemini, and DeepSeek (()). Moreover, numerous efforts have been made to address the inherent limitations of RoPE (()). However, these approaches often rely on augmenting the original RoPE formulation with additional components or interpolation strategies to mitigate issues related to length extrapolation and representation learning, rather than revisiting and revising the core kernel design of RoPE itself. To directly tackle the distance noise problem illustrated in figure 1, we draw inspiration from Lorentz transformations (Hall, 2000). Building on the observation that RoPE can be viewed as a specific case within the broader family of Lorentz transformations, we propose a novel positional encoding scheme. Specifically, we introduce Hyperbolic Rotary Positional Encoding (HoPE), which leverages hyperbolic sine and cosine functions to rotate query and key vectors. The resulting formulation ensures that attention weights decay monotonically with increasing token distance, thereby providing a more stable and robust representation for long-range dependencies.

To validate the effectiveness of HoPE, we conducted extensive experiments in various tasks and datasets. We perform "train short, test extended" perplexity evaluations to assess the model's ability to generalize to sequences longer than those seen during training. Additionally, we evaluated our model on long-text benchmarks to test its performance on tasks requiring the processing of extended sequences. The results demonstrate that HoPE outperforms existing positional encoding methods, achieving lower perplexity and better performance on long-text tasks, thereby confirming the superiority of our approach.

The main contributions of this work can be summarized as follows.

- We revisit classical positional encoding ap-

proaches in Transformers and highlight challenges in length extrapolation, distance-based attention biases, and oscillatory position representations.

- We propose a novel Lorentz rotation framework based on hyperbolic sine and cosine functions, yielding the *Hyperbolic Rotary Positional Encoding* (HoPE) that addresses the oscillatory limitation of RoPE.
- Through extensive experiments on perplexity metrics and long-sequence benchmarks, we empirically demonstrate the superiority of HoPE in short-to-long generalization and overall positional representation quality.

2 Preliminaries

2.1 Relative position encoding

Let $\mathbb{S}_N = \{w_i\}_{i=1}^N$ be a sequence of N input tokens with w_i being the i^{th} element. The corresponding word embedding of \mathbb{S}_N is indicated as $\mathbb{E}_N = \{\mathbf{x}_i\}_{i=1}^N$, where $\mathbf{x}_i \in \mathbb{R}^d$ is the d -dimensional word embedding vector of token w_i without position information.

Relative position encoding aims to represent the relative positional relationships between pairs of tokens. In the context of the attention mechanism, this can be expressed as:

$$\langle f_q(x_m, m), f_k(x_n, n) \rangle = g(x_m, x_n, m - n), \quad (1)$$

where f_q and f_k denote the linear transformations in the attention mechanism, and $\langle \cdot, \cdot \rangle$ represents the inner product operation. The core objective of relative positional encoding is to model the relative positional information between tokens at different positions as accurately as possible while satisfying the given formulation.

2.2 Lorentz Group and Lorentz Transformations

The Lorentz group is a fundamental concept in theoretical physics, representing the Minkowski spacetime symmetry group in special relativity. It encompasses all linear transformations that preserve the spacetime interval between events, ensuring that the laws of physics remain invariant across different inertial frames. The group is continuous and non-compact, characterized by six parameters corresponding to its generators.

The generators of the Lorentz group can be categorized into two types: rotations and boosts. Rotations correspond to transformations that change the spatial orientation of a reference frame without altering its state of motion. At the same time, boosts relate to changes in the inertial frame's velocity along a particular spatial direction.

2.2.1 Finite Rotations in Minkowski Space

Finite rotations around the principal axes in Minkowski spacetime can be represented using specific rotation matrices. These rotations are analogous to those in three-dimensional Euclidean space but are extended to four-dimensional spacetime, preserving the spacetime interval. The rotation matrices for finite rotations about the x , y , and z axes are given by:

$$\begin{aligned} R_x(\theta) &= \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \cos \theta & -\sin \theta \\ 0 & 0 & \sin \theta & \cos \theta \end{pmatrix}, \\ R_y(\psi) &= \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos \theta & 0 & \sin \theta \\ 0 & 0 & 1 & 0 \\ 0 & -\sin \theta & 0 & \cos \theta \end{pmatrix}, \\ R_z(\theta) &= \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos \theta & -\sin \theta & 0 \\ 0 & \sin \theta & \cos \theta & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}. \end{aligned} \quad (2)$$

These matrices act on four-dimensional vectors $(ct, x, y, z)^\top$ and represent rotations in the spatial components while leaving the temporal component unchanged. Here, θ is the rotation angles about the x , y , and z axes, respectively.

2.2.2 Lorentz Boosts

In addition to rotations, the Lorentz group includes boosts, which are transformations between reference frames moving at constant velocities relative to each other. Boosts alter both spatial and temporal components of vectors to preserve the spacetime interval. A standard boost in the x -direction is represented by:

$$B_x(\eta) = \begin{pmatrix} \cosh \eta & -\sinh \eta & 0 & 0 \\ -\sinh \eta & \cosh \eta & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad (3)$$

where η is the rapidity parameter, related to the relative velocity v between frames by $\tanh \eta = v/c$, with c denoting the speed of light. The hyperbolic functions $\cosh \eta$ and $\sinh \eta$ ensure that the spacetime interval remains invariant under the

transformation. Similarly, boosts along the y and z axes are represented by corresponding matrices:

$$\begin{aligned} B_y(\eta) &= \begin{pmatrix} \cosh \eta & 0 & -\sinh \eta & 0 \\ 0 & 1 & 0 & 0 \\ -\sinh \eta & 0 & \cosh \eta & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \\ B_z(\eta) &= \begin{pmatrix} \cosh \eta & 0 & 0 & -\sinh \eta \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ -\sinh \eta & 0 & 0 & \cosh \eta \end{pmatrix}. \end{aligned} \quad (4)$$

3 HoPE

In this section, we introduce **Hyperbolic Rotary Position Encoding (HoPE)**, a positional encoding method formulated from a new perspective. HoPE extends RoPE into hyperbolic space by utilizing Lorentz transformations. This approach leverages the mathematical properties of the Lorentz group to capture complex relative positional relationships within sequences more effectively.

3.1 Hyperbolic Rotary Position Encoding

From the perspective of Lorentz group theory, we draw upon the concept of the *boost generator* in Lorentz transformations to introduce a method of hyperbolic rotation based on hyperbolic trigonometric functions. This approach addresses issues arising from the periodicity inherent in conventional trigonometric functions, such as potential noise due to their cyclic nature. We define the hyperbolic rotation matrix as follows:

$$B(\theta, m) = \begin{pmatrix} \cosh m\theta & \sinh m\theta \\ \sinh m\theta & \cosh m\theta \end{pmatrix} \quad (5)$$

Where $\sinh(\cdot)$ and $\cosh(\cdot)$ represent the hyperbolic sine and cosine functions, respectively. The parameter θ denotes the rotation angle, while it is a scaling factor that adjusts the transformation's magnitude.

Like RoPE, we apply rotations of $m\theta$ and $-m\theta$ to the query (q) and key (k) vectors at position m , respectively. Although RoPE may appear to apply identical transformations to both q and k , the practical effect, due to the matrix transposition that occurs during the computation of the attention mechanism, results in distinct effective rotations of $m\theta$ for q and $-m\theta$ for k . Building upon this concept, we define the rotation matrix for the keys as follows:

$$B'(\theta, m) = \begin{pmatrix} \cosh m\theta & -\sinh m\theta \\ -\sinh m\theta & \cosh m\theta \end{pmatrix} \quad (6)$$

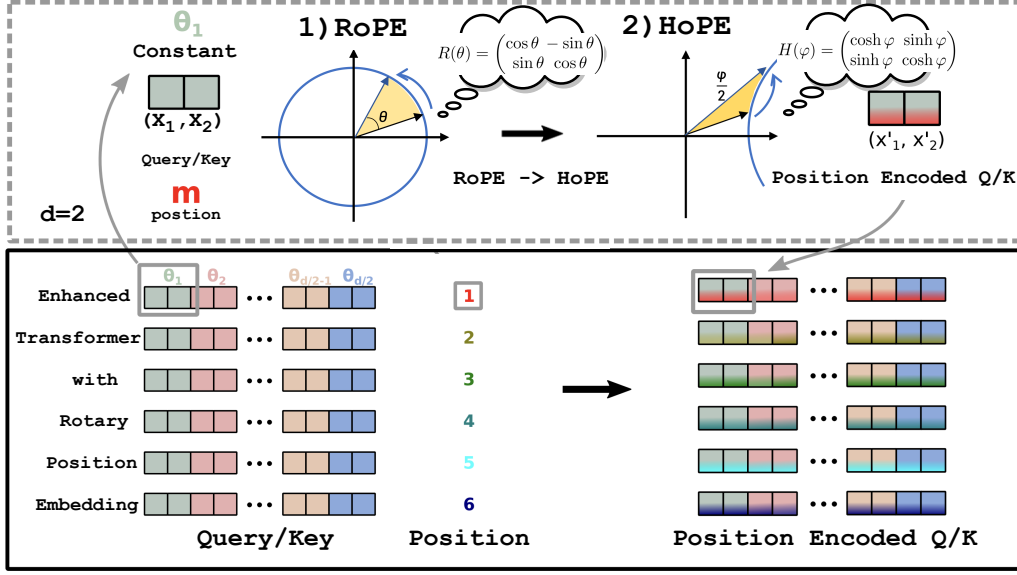


Figure 2: Implementation of Hyperbolic Rotary Position Embedding.

Based on the above equation, the representations of queries and keys along two specific dimensions in self-attention can be expressed as:

$$f_q(\mathbf{x}_m, m) = B(\theta, m)W_{proj}^q \mathbf{x}_m \quad (7)$$

$$f_k(\mathbf{x}_m, m) = B'(\theta, m)W_{proj}^k \mathbf{x}_m \quad (8)$$

Here, W_{proj}^q and W_{proj}^k represent the projection layer weights for queries and keys, respectively, and \mathbf{x}_m denotes the original token embedding at position m . This formalism ensures that the learned positional representations are distinctly handled for queries and keys, leveraging the unique properties of rotational transformations in capturing positional information.

As a result of the action of the attention mechanism, the final form of the dot product is obtained as follows:

$$\begin{aligned} g(\mathbf{x}_m, \mathbf{x}_n, n - m) &= \langle f_q(\mathbf{x}_m, m), f_k(\mathbf{x}_n, n) \rangle \\ &= e^{-(m-n)\theta'} \begin{pmatrix} q_m^{(1)} & q_m^{(2)} \end{pmatrix} B(\theta, m-n) \begin{pmatrix} k_n^{(1)} \\ k_n^{(2)} \end{pmatrix} \end{aligned} \quad (9)$$

However, we find that such transformations alone do not satisfy the assumptions of positional encoding. This is because, unlike RoPE, the Boost generator is not an orthogonal matrix. Furthermore, due to the monotonicity of hyperbolic trigonometric functions, the difference $m - n$ increases the dot product of q and k . As $m - n$ increases, the calculated attention weight of q and k also increases.

This contradicts the assumption of positional encoding, which posits that tokens closer to each other should be assigned higher attention weights.

To address this issue, we introduce a penalty coefficient $e^{\pm m\theta'}$, where θ' is a learnable or predefined parameter, to modulate the positional impact on the dot product. Specifically, the penalty ensures that as the positional difference $m - n$ increases, the dot product of q and k decreases, thereby enforcing the intended behaviour of the attention mechanism. This design ensures that tokens closer in position are prioritized with higher attention weights, enhancing the model's ability to capture local context and long-range dependencies accurately.

The modified query and key representations, which incorporate this penalty coefficient, are defined as follows:

$$f_q(\mathbf{x}_m, m) = e^{-m\theta'} B(\theta, m)W_{proj}^q \mathbf{x}_m, \quad (10)$$

$$f_k(\mathbf{x}_m, m) = e^{m\theta'} B'(\theta, m)W_{proj}^k \mathbf{x}_m, \quad (11)$$

Using $e^{\pm m\theta'}$ ensures a decaying or amplifying effect on the positional encoding component, countering the undesired monotonic increase and aligning the attention mechanism with the theoretical assumptions of positional encodings.

3.2 Theoretical Analysis

3.2.1 Long-range Decay Property

For dimension-pair $(2i, 2i+1)$, consider the asymptotic behavior:

$$\lim_{|m-n| \rightarrow \infty} e^{-|m-n|\theta'} \cosh(|m-n|\theta_i) \propto e^{-|m-n|(\theta' - \theta_i)} \quad (12)$$

When $\theta' > \theta_i, \forall i$, the attention weights exhibit exponential decay concerning positional distance. This satisfies the locality prior while maintaining controlled long-range interaction capability.

3.2.2 Positional Information Capacity

HoPE preserves RoPE’s theoretical advantages in positional discrimination:

Positional Discrimination Capacity. For any relative position $r \in \mathbb{Z}$ and query vector q , \exists key vector k such that:

$$\operatorname{argmax}_s (\langle f_q(m), f_k(m+s) \rangle) = r \quad (13)$$

The hyperbolic rotation creates position-dependent orientation in the embedding space. For target position r , construct k such that $W_k x_n$ aligns with $e^{r\theta'} R'(-\theta, r) W_q x_m$ in the rotated space. This attribute ensures that HoPE maintains the capability to focus on tokens at significant positions while introducing a controllable distance decay.

3.2.3 Generalization to Higher Dimensions

For d -dimensional embeddings (d even), we implement block-diagonal transformations:

$$R_{\Theta, m}^d = \bigoplus_{i=0}^{d/2-1} R(\theta_i, m) \quad (14)$$

Each 2D subspace receives independent rotation parameters θ_i , with global damping controlled by $\theta' > \max_i \theta_i$. The parameterization enables hierarchical capture of positional relationships. These refinements maintain the original contributions while improving mathematical rigor, clarifying causal relationships between components, and emphasizing the approach’s theoretical underpinnings. The narrative flows better, connecting motivation, implementation, and theoretical analysis.

4 Experiments

4.1 Experimental Setup

We evaluate the HoPE’s positional encoding capability using two primary metrics: perplexity for pre-training and performance on downstream tasks with the SCROLLS benchmark.

Perplexity: This metric quantifies the ability of a language model to predict a sequence of words or tokens, with lower perplexity indicating greater confidence and accuracy in its predictions.

Downstream Task Performance: In natural language processing tasks, downstream performance

does not always correlate directly with perplexity. Consequently, we have chosen to use the SCROLLS(Shaham et al., 2022) benchmark to evaluate the impact of various positional encodings on the performance of the downstream task.

4.2 Perplexity Experiment (PPL)

We test the length extrapolation capability of Transformer-based language models with various positional encoding methods. Following the methodology of (Chi et al., 2022b), we use the Pile dataset (Gao et al., 2020) as the pre-training corpus and evaluate the log perplexity of pre-trained language models in the test sets of PG19 (Rae et al., 2019) and arXiv. We conduct non-overlapping evaluations when computing the perplexity score.

The pre-training sequence length is set to 1024, and we evaluate zero-shot perplexity on sequence lengths [1024, 2048, 3072, 4096, 5120, 6144]. We choose the standard decoder-only Transformer(Touvron et al., 2023) as the base model and compare our HoPE method against other positional encoding methods: RoPE and Alibi For general segmentation purposes, and full stops determine the boundaries ""."" and newline characters "\n" The Transformer-based language model configuration includes 12 layers, a hidden dimension of 768, and 12 attention heads, resulting in approximately 155M parameters.

The results, illustrated in Table 1 and Table 2(Best performing results are highlighted in bold), show that our HoPE method consistently outperforms RoPE on sequences longer than the training length. While Alibi achieves the lowest perplexity on the arXiv dataset, as noted in (Chen et al., 2024; Peng et al., 2023), this phenomenon can be attributed to two main factors:

1) The nature of the training corpus: Most texts in our pre-training datasets predominantly exhibit short-distance dependencies, meaning that accurate token prediction primarily relies on information from nearby contexts rather than long-range dependencies.

2) Alibi’s architectural advantage in this scenario: Its linear attention decay mechanism naturally emphasizes local context while attenuating long-distance information. This characteristic aligns well with the short-distance dependency pattern in our training data, resulting in stable perplexity scores even as sequence length increases.

Integrating Hope with interpolation strategies. Recent advancements in the extrapolation of con-

Method	1024	2048	3072	4096	5120	6144
RoPE	12.82	25.80	56.28	88.59	116.63	144.13
Alibi	11.95	25.11	52.54	79.04	107.59	132.80
HoPE	13.35	16.46	35.07	60.03	85.94	110.02
Bipe-RoPE	13.74	14.49	25.05	40.50	54.47	66.64
Bipe-Alibi	11.95	29.06	63.19	91.86	118.54	142.05
Bipe-HoPE	13.71	14.47	24.00	38.78	52.84	65.34

Table 1: Perplexity Performance Comparison on PG19

Method	1024	2048	3072	4096	5120	6144
RoPE	4.81	12.11	36.84	62.78	98.21	132.63
Alibi	4.82	4.89	4.93	4.98	4.95	5.01
HoPE	4.78	8.90	22.28	40.10	60.48	82.04
Bipe-RoPE	4.74	5.73	12.84	20.05	30.84	40.58
Bipe-Alibi	4.82	4.88	4.94	4.97	4.92	4.98
Bipe-HoPE	4.83	5.12	12.20	20.01	28.85	38.34

Table 2: Perplexity Performance Comparison on arXiv

text length involve interpolating language models based on relative positional encoding, utilizing segmented position encoding techniques (Golovneva et al., 2024). To investigate further the performance differences between our HoPE method and other relative positional encodings post-fine-tuning, we employ BiPE (He et al., 2024) to fine-tune language models pre-trained on the PG19 and arXiv datasets. We then evaluated their performance on downstream tasks. The results are presented in Table 1 and Table 2. Similarly to the observations before fine-tuning, BiPE-HoPE outperforms BiPE-RoPE on longer sequences after fine-tuning.

4.3 Fine-Tuning Experiment

To evaluate the model’s performance in understanding extended contexts, following (Ainslie et al., 2023), we further fine-tune the pre-trained checkpoints on the SCROLLS benchmark SCROLLS consists of seven distinct datasets covering various tasks. We employ three evaluation metrics for different tasks: RGL score (ROUGE-L), unigram overlap (F1), and exact match (EM). We fine-tune pre-trained models using a sequence length of 8192 and select the last model checkpoint on the validation set for final evaluation.

As shown in Table 3, although Alibi achieved the best extrapolation performance on the PPL task using the arXiv dataset, HOPE emerged as the superior positional encoding when fine-tuning with

a sequence length of 8192 for downstream tasks. Specifically, HOPE outperformed other positional encodings in four out of seven tasks. In the NarrativeQA task, HOPE scored 1.57 points higher in Rouge-L compared to RoPE, and in the QMSum task, it scored 2.33 points higher than Alibi.

4.4 Ablation Studies

To further analyze HoPE’s effectiveness, we conduct ablation studies that examine the impact of individual components. These studies provide deeper insight into the aspects of HoPE that contribute significantly to its overall performance. The relevant sections of this paper provide Detailed results and discussions of these ablation studies.

Scaling factor is important: We investigated the impact of scaling factors on overall positional encoding within perplexity experiments conducted on arXiv. Specifically, we fixed the hyperbolic rotation angle and observed the effects of different scaling factors on the perplexity of the model.

In these experiments, the scaling factor was defined as the proportion used to adjust the magnitude of positional encodings. Our experimental results indicate that different scaling factors significantly affect the model’s perplexity with the rotation angle held constant. As shown in Figure 3:

Smaller Scaling Factors: When the scaling factor is small, the amplitude of positional encodings decreases, making it difficult for the model to capture

	QAS	CNLI	QMS	NQA	SumS	GovR	QuAL
Metric	F1	EM	RGL	F1	RGL	RGL	EM
Median length	5472	2148	14197	57829	9046	8841	7171
Sinusoidal	9.2	58.9	16.89	6.08	12.74	14.47	1.9
Randomized RoPE	13.02	69.37	16.31	6.89	13.45	16.94	12.8
RoPE	12.98	69.43	16.03	7.57	13.69	15.55	0.68
Alibi	14.17	65.41	14.75	6.73	13.04	18.83	0.87
HoPE	14.52	68.91	17.08	9.14	13.35	19.34	0.45

Table 3: Performance comparison on SCROLLS benchmark. Abbreviations for dataset names: Qasper (Qas), ContractNLI (CNLI), QMSum (QMS), NarrativeQA (NQA), SummScreenFD (SumS), GovReport (GovR), and QuALITY (QuAL). Best performing results are highlighted in **bold**.

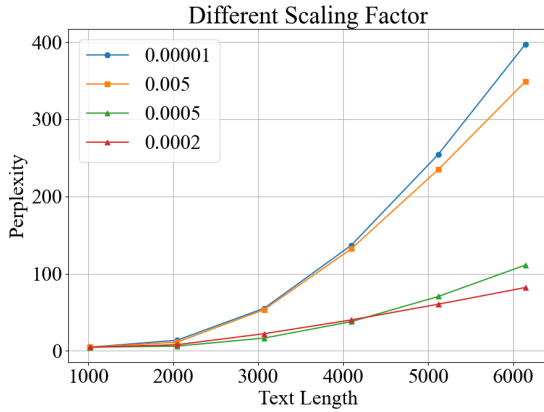


Figure 3: Ablation Experiment

long-range dependencies within sequences, resulting in higher perplexity.

Moderate Scaling Factors: Moderate scaling factors strike a balance by maintaining positional information while avoiding noise amplification or unnecessary details, typically leading to lower perplexity.

Larger Scaling Factors: Substantial scaling factors can amplify noise or other non-ideal characteristics in positional encodings, deteriorating model performance and resulting in higher perplexity.

4.5 Further Analysis from the Attention Weight Perspective

To further analyze the capabilities of various positional encodings, we investigated their impact on attention weights before activation functions are applied. Specifically, we initialized two fixed vectors, q and k , to simulate the effect of different positional encodings on computing attention weights when these vectors are placed at varying positions. The results are illustrated in Figure 4. All positional encodings exhibit a decay characteristic with in-

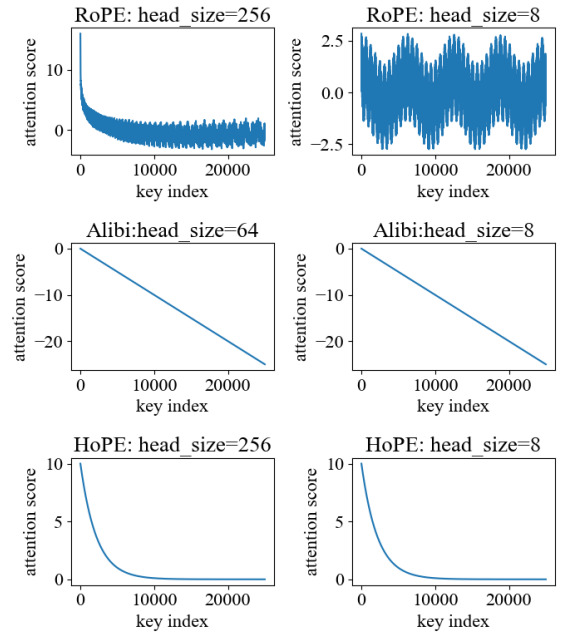


Figure 4: Attention Weight Values under Different Positional Encodings

creasing relative distance. However, during this decay process, Rope demonstrates fluctuation issues due to the periodic nature of trigonometric functions, leading to localized decreases in higher frequency dimensions. In contrast, the decline observed in Alibi is not smooth, suggesting that it could be more accurately described as an attention bias rather than a positional encoding. In particular, the Hope was meticulously designed to overcome these challenges, resulting in smoother and more stable performance across different relative distances.

Furthermore, instead of assigning fixed values to these vectors, each q and k vector was independently initialized using a Gaussian distribution,

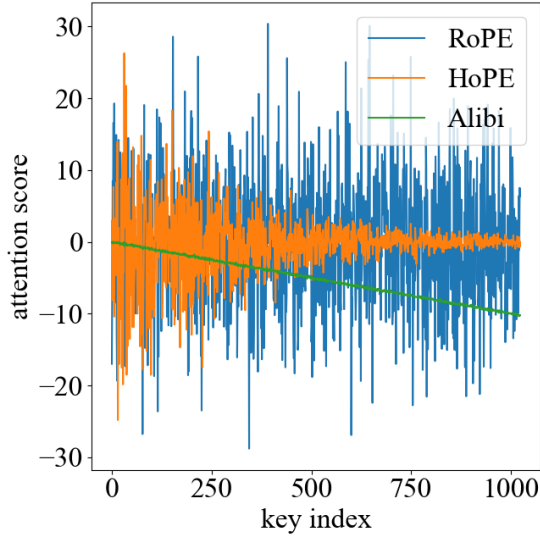


Figure 5: Attention Weight Values under Different Positional Encodings

which was more reflective of real-world scenarios. The resulting attention weights, derived from the inner product of q and k , were plotted in Fig. 5. Our findings reveal that RoPE modifies the inner product between q and k vectors by adjusting their angular relationships. Although RoPE effectively captures long-range dependencies when q and k vectors are aligned in a fixed direction, this alignment does not consistently apply across all dimensions of high-dimensional vectors when q and k are randomly initialized. Specifically, RoPE’s ability to preserve locality among closely related dimensions appears limited; random initialization of q and k (implying random angles between them) diminishes RoPE’s capacity to maintain local relevance, resembling situations without positional encoding (NoPE). Furthermore, initializing vectors with a Gaussian distribution suggests that specific tokens should attract more attention than others, with attention decreasing as distance increases a behavior consistent with our expectations. However, one drawback observed with AliBi is that distant but relevant tokens sometimes receive less attention than closer, irrelevant ones.

5 Related work

Existing positional encoding mechanisms can be classified into absolute and relative positional encodings, each having different trade-offs in length extrapolation and dependency modeling. Absolute Positional Encodings, pioneered by (Vaswani et al.,

2023) using fixed sinusoidal patterns or learnable embeddings, are simple but struggle with length extrapolation due to their reliance on predefined positional indices (Devlin et al., 2019). Studies (Wang and Chen, 2020) have shown that while sinusoidal embeddings can implicitly capture positional relationships, they fail to generalize effectively beyond training sequence lengths. Relative Positional Encodings, introduced by (Shaw et al., 2018), model pairwise token distances through additive biases in attention scores, offering greater flexibility but encountering scalability challenges with long-range dependencies. Alibi (Press et al., 2022) enhanced extrapolation capability through a distance-decaying linear attention bias, but its heuristic design lacks guarantees for monotonic decay, leading to suboptimal performance on extremely long sequences. Rotary Positional Encodings (RoPE) (Su et al., 2023), implementing rotation-based positional encoding through trigonometric transformations, theoretically maintains relative positional relationships across different sequence lengths. However, RoPE exhibits oscillatory attention patterns due to its trigonometric periodicity, which can destabilize long-distance dependency modeling (Barbero et al., 2024).

Hyperbolic space, regarded as the continuous analogue of discrete trees (Krioukov et al., 2010), offers inherent advantages for modeling data with implicit or explicit tree-like structures, including hierarchical organizations and power-law distributions (Adcock et al., 2013; Zhou et al., 2022). Recent advances in representation learning have extensively demonstrated the superiority of hyperbolic geometry through multiple perspectives: low-distortion embeddings (Sarkar, 2011), reduced generalization error (Suzuki et al., 2021a,b), and superior empirical performance (Chami et al., 2019; Yang et al., 2022). These advantages have been successfully leveraged across diverse research domains and downstream applications (Peng et al., 2021; Song et al., 2023c, 2022a, 2023b; Mettes et al., 2023; Song et al., 2023a, 2022b; Yang et al., 2021), encompassing graph learning, computer vision, and natural language processing.

6 Conclusion

We propose HoPE to address the oscillatory attention patterns in RoPE that limit long-range dependency modeling. Based on Lorentz transformations, HoPE replaces trigonometric rotations with

hyperbolic functions to ensure monotonic attention decay with increasing token distances. Theoretical analysis validates that HoPE’s geometric design naturally achieves distance-aware attention. Extensive experiments demonstrate HoPE’s advantages over existing methods. It exhibits superior length extrapolation capabilities in "train short, test long" scenarios and achieves state-of-the-art performance on long-text tasks. These results verify that HoPE effectively maintains stable positional representations while preserving the essential rotational invariance of Transformer attention.

7 Limitations

In this paper, we propose HoPE to address limitations in current positional encoding methods. By leveraging Lorentz transformations and hyperbolic functions, HoPE yields more stable position representations. Theoretical analysis and extensive experiments demonstrate its advantages in capturing long-range dependencies and enabling length extrapolation. However, HoPE still faces challenges. First, while it excels in text-only tasks, its performance in multimodal scenarios (where text, audio, and visual inputs must be jointly modelled) remains unverified. Second, the method’s effectiveness hinges on careful tuning of the damping coefficient θ' . Suboptimal choices can degrade performance, especially for tasks with varying positional sensitivity requirements.

References

- Aaron B Adcock, Blair D Sullivan, and Michael W Mahoney. 2013. Tree-like structure in large social and information networks. In *IEEE International Conference on Data Mining*, pages 1–10. IEEE.
- Joshua Ainslie, Tao Lei, Michiel de Jong, Santiago Ontañón, Siddhartha Brahma, Yury Zemlyanskiy, David Uthus, Mandy Guo, James Lee-Thorp, Yi Tay, Yun-Hsuan Sung, and Sumit Sanghai. 2023. *Colt5: Faster long-range transformers with conditional computation*. *Preprint*, arXiv:2303.09752.
- Cem Anil, Yuhuai Wu, Anders Andreassen, Aitor Lewkowycz, Vedant Misra, Vinay Ramasesh, Ambrose Slone, Guy Gur-Ari, Ethan Dyer, and Behnam Neyshabur. 2022. *Exploring length generalization in large language models*. *Preprint*, arXiv:2207.04901.
- Federico Barbero, Alex Vitvitskiy, Christos Perivolaropoulos, Razvan Pascanu, and Petar Veličković. 2024. *Round and round we go! what makes rotary positional encodings useful?* *Preprint*, arXiv:2410.06205.
- Ines Chami, Zhitao Ying, Christopher Ré, and Jure Leskovec. 2019. Hyperbolic graph convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 4868–4879.
- Guanzheng Chen, Xin Li, Zaiqiao Meng, Shangsong Liang, and Lidong Bing. 2024. *Clex: Continuous length extrapolation for large language models*. *Preprint*, arXiv:2310.16450.
- Mingda Chen, Zewei Chu, Sam Wiseman, and Kevin Gimpel. 2022. *Summscreen: A dataset for abstractive screenplay summarization*. *Preprint*, arXiv:2104.07091.
- Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. 2023. *Extending context window of large language models via positional interpolation*. *Preprint*, arXiv:2306.15595.
- Ta-Chung Chi, Ting-Han Fan, Peter J Ramadge, and Alexander Rudnicky. 2022a. Kerple: Kernelized relative positional embedding for length extrapolation. *Advances in Neural Information Processing Systems*, 35:8386–8399.
- Ta-Chung Chi, Ting-Han Fan, Peter J. Ramadge, and Alexander I. Rudnicky. 2022b. *Kerple: Kernelized relative positional embedding for length extrapolation*. *Preprint*, arXiv:2205.09921.
- Ta-Chung Chi, Ting-Han Fan, Peter J Ramadge, et al. 2023. Dissecting transformer length extrapolation via the lens of receptive field analysis. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.
- Jishnu Ray Chowdhury and Cornelia Caragea. 2023. *Monotonic location attention for length generalization*. *Preprint*, arXiv:2305.20019.
- Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A. Smith, and Matt Gardner. 2021. *A dataset of information-seeking questions and answers anchored in research papers*. *Preprint*, arXiv:2105.03011.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. *Bert: Pre-training of deep bidirectional transformers for language understanding*. *Preprint*, arXiv:1810.04805.
- Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020. *The pile: An 800gb dataset of diverse text for language modeling*. *Preprint*, arXiv:2101.00027.
- Olga Golovneva, Tianlu Wang, Jason Weston, and Sainbayar Sukhbaatar. 2024. *Contextual position encoding: Learning to count what’s important*. *Preprint*, arXiv:2405.18719.
- Brian C. Hall. 2000. *An elementary introduction to groups and representations*. *Preprint*, arXiv:math-ph/0005032.

- Adi Haviv, Ori Ram, Ofir Press, Peter Izsak, and Omer Levy. 2022. [Transformer language models without positional encodings still learn positional information](#). *Preprint*, arXiv:2203.16634.
- Zhenyu He, Guhao Feng, Shengjie Luo, Kai Yang, Liwei Wang, Jingjing Xu, Zhi Zhang, Hongxia Yang, and Di He. 2024. [Two stones hit one bird: Bilevel positional encoding for better length extrapolation](#). *Preprint*, arXiv:2401.16421.
- Luyang Huang, Shuyang Cao, Nikolaus Parulian, Heng Ji, and Lu Wang. 2021. [Efficient attentions for long document summarization](#). *Preprint*, arXiv:2104.02112.
- Amirhossein Kazemnejad, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Payel Das, and Siva Reddy. 2023. [The impact of positional encoding on length generalization in transformers](#). *Preprint*, arXiv:2305.19466.
- Yuta Koreeda and Christopher D. Manning. 2021. [Contractnli: A dataset for document-level natural language inference for contracts](#). *Preprint*, arXiv:2110.01799.
- Tomáš Kočiský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2017. [The narrativeqa reading comprehension challenge](#). *Preprint*, arXiv:1712.07040.
- Dmitri Krioukov, Fragkiskos Papadopoulos, Maksim Kitsak, Amin Vahdat, and Marián Boguná. 2010. Hyperbolic geometry of complex networks. *Physical Review E*, 82(3):036106.
- Shanda Li, Chong You, Guru Guruganesh, Joshua Ainslie, Santiago Ontanon, Manzil Zaheer, Sumit Sanghai, Yiming Yang, Sanjiv Kumar, and Srinadh Bhojanapalli. 2024. [Functional interpolation for relative positions improves long context transformers](#). *Preprint*, arXiv:2310.04418.
- Pascal Mettes, Mina Ghadimi Atigh, Martin Keller-Ressel, Jeffrey Gu, and Serena Yeung. 2023. Hyperbolic deep learning in computer vision: A survey. *arXiv preprint arXiv:2305.06611*.
- Richard Yuanzhe Pang, Alicia Parrish, Nitish Joshi, Nikita Nangia, Jason Phang, Angelica Chen, Vishakh Padmakumar, Johnny Ma, Jana Thompson, He He, and Samuel R. Bowman. 2022. [Quality: Question answering with long input texts, yes!](#) *Preprint*, arXiv:2112.08608.
- Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. 2023. [Yarn: Efficient context window extension of large language models](#). *Preprint*, arXiv:2309.00071.
- Wei Peng, Tuomas Varanka, Abdelrahman Mostafa, Henglin Shi, and Guoying Zhao. 2021. Hyperbolic deep neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Ofir Press, Noah A. Smith, and Mike Lewis. 2022. [Train short, test long: Attention with linear bi-ases enables input length extrapolation](#). *Preprint*, arXiv:2108.12409.
- Jack W. Rae, Anna Potapenko, Siddhant M. Jayakumar, and Timothy P. Lillicrap. 2019. [Compressive transformers for long-range sequence modelling](#). *Preprint*, arXiv:1911.05507.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. [Exploring the limits of transfer learning with a unified text-to-text transformer](#). *Preprint*, arXiv:1910.10683.
- Anian Ruoss, Grégoire Delétang, Tim Genewein, Jordi Grau-Moya, Róbert Csordás, Mehdi Bennani, Shane Legg, and Joel Veness. 2023. [Randomized positional encodings boost length generalization of transformers](#). *Preprint*, arXiv:2305.16843.
- Rik Sarkar. 2011. Low distortion delaunay embedding of trees in hyperbolic plane. In *International Symposium on Graph Drawing*, pages 355–366. Springer.
- Uri Shaham, Elad Segal, Maor Ivgi, Avia Efrat, Ori Yoran, Adi Haviv, Ankit Gupta, Wenhan Xiong, Mor Geva, Jonathan Berant, and Omer Levy. 2022. [Scrolls: Standardized comparison over long language sequences](#). *Preprint*, arXiv:2201.03533.
- Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. 2018. [Self-attention with relative position representations](#). *Preprint*, arXiv:1803.02155.
- Mingyang Song, Yi Feng, and Liping Jing. 2022a. [Hyperbolic relevance matching for neural keyphrase extraction](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 5710–5720. Association for Computational Linguistics.
- Mingyang Song, Yi Feng, and Liping Jing. 2022b. [A preliminary exploration of extractive multi-document summarization in hyperbolic space](#). In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management, Atlanta, GA, USA, October 17-21, 2022*, pages 4505–4509. ACM.
- Mingyang Song, Yi Feng, and Liping Jing. 2023a. [Hisum: Hyperbolic interaction model for extractive multi-document summarization](#). In *Proceedings of the ACM Web Conference 2023, WWW 2023, Austin, TX, USA, 30 April 2023 - 4 May 2023*, pages 1427–1436. ACM.
- Mingyang Song, Huafeng Liu, Yi Feng, and Liping Jing. 2023b. [Improving embedding-based unsupervised keyphrase extraction by incorporating structural information](#). In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 1041–1048. Association for Computational Linguistics.

Mingyang Song, Huafeng Liu, and Liping Jing. 2023c. [Hyperrank: Hyperbolic ranking model for unsupervised keyphrase extraction](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 16070–16080. Association for Computational Linguistics.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. 2023. [Roformer: Enhanced transformer with rotary position embedding](#). *Preprint*, arXiv:2104.09864.

Yutao Sun, Li Dong, Barun Patra, Shuming Ma, Shao-han Huang, Alon Benhaim, Vishrav Chaudhary, Xia Song, and Furu Wei. 2022. [A length-extrapolatable transformer](#). *Preprint*, arXiv:2212.10554.

Atsushi Suzuki, Atsushi Nitanda, Jing Wang, Linchuan Xu, Kenji Yamanishi, and Marc Cavazza. 2021a. Generalization error bound for hyperbolic ordinal embedding. In *International Conference on Machine Learning*, pages 10011–10021. PMLR.

Atsushi Suzuki, Atsushi Nitanda, Linchuan Xu, Kenji Yamanishi, Marc Cavazza, et al. 2021b. Generalization bounds for graph embedding using negative sampling: Linear vs hyperbolic. *Advances in Neural Information Processing Systems*, 34:1243–1255.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. [Llama 2: Open foundation and fine-tuned chat models](#). *Preprint*, arXiv:2307.09288.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. [Attention is all you need](#). *Preprint*, arXiv:1706.03762.

Yu-An Wang and Yun-Nung Chen. 2020. [What do position embeddings learn? an empirical study of pre-trained language model positional encoding](#). *Preprint*, arXiv:2010.04903.

Wenhan Xiong, Jingyu Liu, Igor Molybog, Hejia Zhang, Prajjwal Bhargava, Rui Hou, Louis Martin, Rashi

Rungta, Karthik Abinav Sankararaman, Barlas Oguz, Madian Khabsa, Han Fang, Yashar Mehdad, Sharan Narang, Kshitiz Malik, Angela Fan, Shruti Bhosale, Sergey Edunov, Mike Lewis, Sinong Wang, and Hao Ma. 2023. [Effective long-context scaling of foundation models](#). *Preprint*, arXiv:2309.16039.

Haoran Yang, Hongxu Chen, Lin Li, Philip S Yu, and Guandong Xu. 2021. Hyper meta-path contrastive learning for multi-behavior recommendation. *arXiv preprint arXiv:2109.02859*.

Menglin Yang, Zhihao Li, Min Zhou, Jiahong Liu, and Irwin King. 2022. HICF: Hyperbolic informative collaborative filtering. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2212–2221.

Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan Awadallah, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, and Dragomir Radev. 2021. [Qmsum: A new benchmark for query-based multi-domain meeting summarization](#). *Preprint*, arXiv:2104.05938.

Min Zhou, Bisheng Li, Menglin Yang, and Lujia Pan. 2022. Telegraph: A benchmark dataset for hierarchical link prediction. *arXiv preprint arXiv:2204.07703*.

8 Appendix

8.1 Proof

Theorem 3.2.2 (Positional Discrimination Capacity) For any relative position $r \in \mathbb{Z}$ and query vector q , \exists key vector k such that:

$$\operatorname{argmax}_s (\langle f_q(m), f_k(m+s) \rangle) = r \quad (15)$$

Frist, considering the 2-dimensional case, the attention weight can be calculated as:

$$V(q_m, k_n) = \mathbf{q}_m^T B(\theta, m) B'(\theta, n) \mathbf{k}_n \quad (16)$$

if

$$\operatorname{argmax}_s (\langle f_q(m), f_k(m+s) \rangle) = t \neq r \quad (17)$$

which means

$$V(q_m, k_t) > V(q_m, k_r) \quad (18)$$

such that

$$\mathbf{q}_m^T B(\theta, m) B'(\theta, t) \mathbf{k}_t > \mathbf{q}_m^T B(\theta, m) B'(\theta, r) \mathbf{k}_r \quad (19)$$

We only need to reconstruct \mathbf{k}_r so that

$$\mathbf{k}_r > \frac{B'(\theta, t)}{B'(\theta, r)} \mathbf{k}_t \quad (20)$$

So, in 2-dimension case, \exists key vector k such that:

$$\operatorname{argmax}_s (\langle f_q(m), f_k(m+s) \rangle) = r \quad (21)$$

When generalized to 2-n dimensions, the theorem maintains its original properties. By constructing k in this way for every 2-dimensional subspace, it ensures that

$$\operatorname{argmax}_s \sum_{k=1}^n \left(\mathbf{q}_m^{(k)} \right)^\top \rho(g_k)^{m-n} \mathbf{k}_n^{(k)} = r \quad (22)$$

8.2 Analysis of RoPE and HoPE

8.2.1 RoPE

For simplicity of notation in this work, we follow the assumption (Barbero et al., 2024) that queries and keys are d dimensional vectors with $d \geq 2$ being an even number. We decompose queries and keys into 2-dimensional chunks $\mathbf{q}_i = \bigoplus_{k=1, \dots, d/2} \mathbf{q}_i^{[k, k+1]} = \bigoplus_{k=1, \dots, d/2} \mathbf{q}_i^{(k)}$, where \bigoplus denotes direct sum (concatenation). In other words, we denote by $\mathbf{q}_i^{(k)} \in \mathbb{R}^2$ the k -th 2-dimensional chunk of the query vector of the i -th token, using analogous notation for the key vectors.

RoPE considers a sequence of angles $G = (g_k = \theta^{-2(k-1)/d} : k = 1, \dots, d/2)^2$, where $g_1 = 1$ is the fastest rotating component at 1 radian per token and $g_{d/2} = \theta^{-(d-2)/d} \approx \theta^{-1}$ the slowest rotating component at approximately $1/\theta$ rotations per token. The parameter θ is called the base wavelength, which by default is 10,000. We denote by $\rho(g_k)$ the matrix form of g_k :

$$\rho(g_k) = \begin{bmatrix} \cos(g_k) & -\sin(g_k) \\ \sin(g_k) & \cos(g_k) \end{bmatrix}, \quad (23)$$

highlighting that $\rho(g_k)$ is a 2-dimensional orthogonal transformation (rotation). One can view $\rho(g_k)$ as a ‘unit rotation’ by g_k radians. The RoPE technique amounts to the construction of a block-diagonal matrix $\mathbf{R}^i = \bigoplus_{k=1, \dots, d/2} \rho(g_k)^i \in \mathbb{R}^{d \times d}$, where each 2×2 block on the diagonal is a rotation by a different frequency of RoPE. The \mathbf{R}^i denotes in fact matrix exponentiation by an integer i , which is the position of \mathbf{x}_i . We can exploit a nice property of rotation matrices, i.e., that $\rho(g_k)^i = \rho(ig_k)$ to avoid the computation of the matrix power. As this matrix is block diagonal, computing $\mathbf{R}_i \mathbf{q}_i$ means that the rotations act only on 2-dimensional chunks of the query (or key), i.e., $\mathbf{R}_i \mathbf{q}_i = \bigoplus_{k=1, \dots, d/2} \rho(ig_k) \mathbf{q}_i^{(k)}$. This leads to the final formulation of k_{RoPE} :

$$\begin{aligned} k_{\text{RoPE}}(\mathbf{q}_i, \mathbf{k}_j) &= (\mathbf{R}^i \mathbf{q}_i)^\top (\mathbf{R}^j \mathbf{k}_j) \\ &= \mathbf{q}_i^\top \mathbf{R}^{j-i} \mathbf{k}_j \\ &= \sum_{k=1, \dots, d/2} \left(\mathbf{q}_i^{(k)} \right)^\top \rho(g_k)^{j-i} \mathbf{k}_j^{(k)} \end{aligned} \quad (24)$$

where we use the fact that $(\rho(g_k)^i)^\top \rho(g_k)^j = \rho(g_k)^{-i} \rho(g_k)^j = \rho(g_k)^{j-i}$. We highlight how the block diagonal structure of \mathbf{R} allows one to decompose the dot product into the sum of dot products of 2-dimensional chunks, with each key vector chunk rotated at a frequency dictated by g_k .

Considering the smallest and largest values of θ in RoPE, denoted as θ_{\min} and θ_{\max} , respectively, and their corresponding wavelengths, λ_{\max} and λ_{\min} . When a sequence is input into the model, it essentially performs an uneven positional encoding on the sequence.

When the sequence length begins to exceed λ_{\min} , dimensions start completing one cycle of rotation. As the sequence length continues to increase, more dimensions reach their rotational cycles. This phenomenon endows RoPE with the ability to extrapolate. When the sequence length increases, some dimensions’ rotational values have been seen during the model’s previous training, providing familiarity for longer sequences. However, this also introduces noise into the training process. As the sequence length increases, the attention weights (attn_weight) on these dimensions may become larger than those for shorter relative positions, confusing the model. This issue is difficult to mitigate by merely adjusting the frequency. The extrapolation capability of RoPE arises from this characteristic. In extreme cases, if the sequence length exceeds the lowest frequency, i.e., the longest wavelength λ_{\max} , RoPE degenerates into no position encoding, failing to provide positional information to the model anymore.

8.2.2 HoPE

By substituting the rotation matrix of the RoPE (Rotary Position Embedding) with a hyperbolic rotation matrix, we characterize the relative information by rotating q by $m\theta$ and k by $-m\theta$. This approach employs hyperbolic matrices to encode the positional information differently, capturing the relative positions between tokens effectively. Consistent with the matrix form mentioned above, we

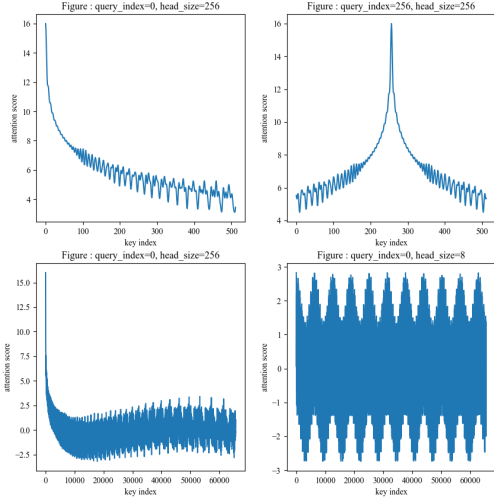


Figure 6: Attention weight decay trend with rope

can derive the following matrix.

$$\rho(g_k) = \begin{bmatrix} \cosh(g_k) & \sinh(g_k) \\ \sinh(g_k) & \cosh(g_k) \end{bmatrix}, \quad (25)$$

However, the rotation matrix of RoPE is an orthogonal matrix, which means it will not change the modulus of any vector:

$$\begin{aligned} & \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix} \begin{bmatrix} q_1 \\ q_2 \end{bmatrix} \\ &= \begin{bmatrix} q_1 \cos(\theta) - q_2 \sin(\theta) \\ q_1 \sin(\theta) + q_2 \cos(\theta) \end{bmatrix} \end{aligned} \quad (26)$$

The length after rotation remains unchanged. However, our hyperbolic rotation is not an orthogonal matrix, and it will change the modulus of the vector.

$$\begin{aligned} & \begin{bmatrix} \cosh(\theta) & \sinh(\theta) \\ \sinh(\theta) & \cosh(\theta) \end{bmatrix} \begin{bmatrix} q_1 \\ q_2 \end{bmatrix} \\ &= \begin{bmatrix} q_1 \cosh(\theta) + q_2 \sinh(\theta) \\ q_1 \sinh(\theta) + q_2 \cosh(\theta) \end{bmatrix} \end{aligned} \quad (27)$$

Original length:

$$\sqrt{q_1^2 + q_2^2} = A \quad (28)$$

Rotated length:

$$\sqrt{(q_1^2 + q_2^2) \cosh(2\theta) + 2q_1 q_2 \sinh(2\theta)} = B \quad (29)$$

To find the relationship between A and B, we got B divided by A

$$\frac{B}{A} = \sqrt{\cosh(2\theta) + \frac{2q_1 q_2 \sinh(2\theta)}{q_1^2 + q_2^2}} \quad (30)$$

According to the inequality related to the binomial theorem: $a^2 + b^2 \geq 2ab$.

$$\begin{aligned} \frac{B}{A} &\leq \sqrt{e^{2\theta}} \\ &= e^\theta \end{aligned} \quad (31)$$

Therefore, to account for the corresponding multiplication of the modulus change, we premultiply the hyperbolic rotation matrix by a penalty coefficient.

$$e^{-\theta} \begin{bmatrix} \cosh(\theta) & \sinh(\theta) \\ \sinh(\theta) & \cosh(\theta) \end{bmatrix} \quad (32)$$

This approach not only reduces the noise during training but also preserves the assumptions of positional encoding.

8.3 Experimental Details

8.3.1 Perplexity Experiment

Model configurations. In this experiment, we train decoder-only Transformer language models with different positional encoding techniques while keeping all the other configurations the same. For RoPE, we follow (Su et al., 2023) to set the hyperparameters in the rotary matrix, respectively. For ALiBi, we follow (Press et al., 2022) to set the slope values in each attention head. For the intra-segment encoding of BiPE, we use the learnable absolute positional encoding. For the inter-segment encoding of BiPE-RoPE, the hyperparameters are kept the same as the original setting. For the inter-segment encoding of our BiPE-ALiBi, the slope values are set to 96 times of the original ALiBi’s setting. Other model configurations are provided in Table 4.

Table 4: Model configurations for length extrapolation.

Layers	12
Attention heads	12
Head dimensions	64
Hidden dimensions	768
FFN dimensions	3072
Model parameters	155M

Training recipes. The next token prediction objective is adopted for language model training. All

models are trained on the Pile dataset with a total sequence length of 1024. The training recipes are shown in Table 5.

Table 5: Training recipes for length extrapolation.

Batch size	256
Total training epochs	1
Dropout	0.0
Weight decay	0.01
Optimizer	AdamW
Learning rate	$1e - 4$

8.3.2 Fine-Tuning Experiment

Fine-tuning on SCROLLS. We fine-tune pre-trained language models with different positional encoding methods on SCROLLS(Shaham et al., 2022). It is a long context benchmark that consists of seven distinct datasets covering different tasks, e.g, Question-Answering (Qasper(Dasigi et al., 2021), NarrativeQA(Kočiský et al., 2017), and QuALITY(Pang et al., 2022)), Natural Language Inference (ContractNLI(Koreeda and Manning, 2021)) and Summarization (QMSum(Zhong et al., 2021), SummScreenFD(Chen et al., 2022), and GovReport(Huang et al., 2021)). All the model configurations are the same as those in Table 4.

Table 6: Finetuning recipes for long context benchmark.

Batch size	64
Total training steps	5000
Dropout	0.0
Weight decay	0.01
Optimizer	AdamW
Learning rate	$1e - 5$

Fine-tuning recipes. We fine-tune models using the next token prediction objective on each task with a sequence length of 8192. The fine-tuning recipes are provided in Table 6

8.4 Lorentz Transformation And Lorentz Group

The general form of the Lorentz transformation can be written as:

$$x'^{\mu} = \Lambda^{\mu}_{\nu} x^{\nu}$$

where Λ^{μ}_{ν} is the Lorentz transformation matrix satisfying the condition:

$$\eta_{\rho\sigma} \Lambda^{\rho}_{\mu} \Lambda^{\sigma}_{\nu} = \eta_{\mu\nu}$$

Algorithm 1: HoPE

Input: q, k: [batch, head, seq, dim]; theta: [dim//2] (per-dimension angle); theta_prime

Output: Modified q and k

for $i \leftarrow 0$ **to** $dim - 2$ **step** 2 **do**

angle \leftarrow pos $\cdot \theta[i/2]$;

$c, s \leftarrow \cosh(\text{angle}), \sinh(\text{angle})$;

rot_q $\leftarrow [c \cdot q[\dots, i] + s \cdot q[\dots, i + 1], s \cdot q[\dots, i] + c \cdot q[\dots, i + 1]]$;

rot_k $\leftarrow [c \cdot k[\dots, i] - s \cdot k[\dots, i + 1], -s \cdot k[\dots, i] + c \cdot k[\dots, i + 1]]$;

$q[\dots, i : i + 2] \leftarrow \exp(-\text{pos} \cdot \theta_{\text{prime}}) \cdot \text{rot_q}$;

$k[\dots, i : i + 2] \leftarrow \exp(\text{pos} \cdot \theta_{\text{prime}}) \cdot \text{rot_k}$;

end

with $\eta_{\mu\nu}$ being the Minkowski metric:

$$\eta_{\mu\nu} = \text{diag}(1, -1, -1, -1)$$

For a specific case of a boost along the x -axis, the transformations are given by:

$$\begin{aligned} t' &= \gamma \left(t - \frac{vx}{c^2} \right) \\ x' &= \gamma(x - vt) \\ y' &= y \\ z' &= z \end{aligned}$$

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

The set of all such transformations forms the Lorentz group, denoted as $O(1, 3)$, which is defined as:

$$O(1, 3) \equiv \underbrace{\{ \Lambda \mid \Lambda \in GL(4, \mathbb{R}), g_{\mu\nu} \Lambda^{\mu}_{\rho} \Lambda^{\nu}_{\sigma} = g_{\rho\sigma} \}}_{\dim O(1,3)=6}$$

This group includes rotations and boosts (velocity transformations) and has six degrees of freedom: three for rotations and three for boosts. It represents the fundamental symmetry group of special relativity, ensuring the invariance of physical laws across different inertial frames.

Essentially, $O(1, 3)$ is the group of linear transformations on Minkowski spacetime that preserve the metric. For more details on metric spaces and index notation, refer to Masaki Notation.