Tree Rotary Positional Encoding for Extreme Length Extrapolation from Scratch

Qiuwu Chen[1]* **Ziteng Huang**[2]* **Shuhai Zhang**[2]* **Zimo Liu**[1]
Linxiao Li[1] **Ying Sun**[1] **Yuchen Li**[1]
Yifan Zhang[1] **Yaofo Chen**[2,3]† **Mingkui Tan**[2]†

[1]AIGCode
[2]South China University of Technology
[3]Pazhou Lab

Abstract

Large language models rely on positional representations for long context modeling, yet are trained on short sequences and evaluated at much longer lengths. Bridging this gap requires positional encodings remaining distinguishable at large distances while generalizing beyond the training length. In this paper, we identify a *structural* limitation of existing rotary-based positional encodings: all positional dimensions are controlled by a *single* global index, which couples training-time positional variation with long-range extrapolation. To address this, we propose Tree Rotary Positional Encoding (TPE), a structured *multi-scale* positional encoding designed for training from scratch that represents absolute positions through compositions of different components, avoiding single-index extrapolation. We further devise an Offset-based Positional Training (OPT) strategy to improve learning interactions across TPE components, enabling better generalization to longer contexts. We theoretically analyze how TPE's multi-scale decomposition enables effective combinatorial position coverage. On a 1.2B model trained from scratch at length 512, TPE reduces 32K perplexity from 4532.21 (RoPE) to 149.96, with comparable inference efficiency. This 64$\times$ extrapolation characterizes next-token likelihood; separately, Passkey retrieval accuracy is 58.4% at 8K and 38.1% at 16K.

1 Introduction

Large language models (LLMs) [1, 2] have demonstrated strong capabilities across a wide range of challenging tasks, including question answering [3, 4, 5], code generation [6, 7, 8], and document understanding [9, 10, 11]. Most models are trained on relatively short sequences but are required to reason over longer contexts in inference, making the design of positional representations crucial for generalization beyond the training context [12, 13, 14].

Encoding positional information that supports such long-range distinguishability is fundamentally challenging. Positional representations must remain sufficiently discriminative as positional differences grow large, so that relative ordering and separation between distant tokens can still be distinguished [13]. At the same time, these representations are learned from models trained on finite context lengths, yet are expected to operate at much longer ranges during inference [15]. Handling this mismatch between training and inference remains a challenge for long-context modeling.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

*Equal contribution. Emails: Qiuwu Chen (chenqiuwu@aigcode.net), Ziteng Huang (goodluckhzt@gmail.com), and Shuhai Zhang (shuhaizhangshz@gmail.com).

†Corresponding authors. Emails: Yaofo Chen (chenyaofo@scut.edu.cn) and Mingkui Tan (mingkuitan@scut.edu.cn).

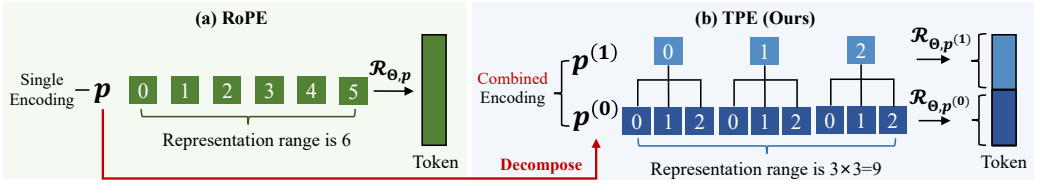


Figure 1: Comparisons between single-index rotary positional encoding (RoPE) and our Tree Rotary Positional Encoding (TPE). RoPE uses a single absolute index p, limiting the representation range within the training context. TPE decomposes positions into multiple components (*e.g.*, $p^{(0)}$,$p^{(1)}$) across scales and composes them to form position representations. This yields a large position space, enabling long-context positions to be represented by combining a small set of learned components.

Most modern LLMs adopt Rotary Positional Encoding (RoPE) [16], which encodes positions by rotating feature subspaces using an absolute positional index. By incorporating relative positional information into attention, RoPE partially addresses the need for long-range distinguishability. However, its reliance on a single global index introduces inherent limitations: low-frequency dimensions vary little within the training context, and extrapolation to longer sequences pushes all dimensions into unseen rotation phases. A recent study [17] mitigates these effects through rescaling or interpolation, but largely retain the same single-index parameterization and its associated constraints.

Motivated by these, we argue that the limitation of existing rotary-based methods lies not in specific frequency choices, but in the use of a single positional index to control all positional dimensions. This design tightly couples positional variation in training with behavior under long-range extrapolation, making it difficult to achieve both strong in-context discrimination and reliable generalization to longer sequences. To break this coupling, positional representations should be structured across multiple scales, so that long-range positions can be expressed through the composition of bounded, reusable components, rather than by extrapolating a single index far beyond its training range.

Based on this insight, we propose **Tree Rotary Positional Encoding (TPE)**, a structured multi-scale rotary encoding that addresses the single-index limitation through bounded radix composition. TPE represents an absolute position as bounded, reusable components across scales (Figure 1), interpreted as a path through a tree. Their composition supports long-range positional distinguishability without extrapolating a single unbounded index or requiring semantic segmentation. This bounded composition, beyond hierarchical dimension splitting alone, defines its positional parameterization. To further improve long-range behavior under finite training lengths, we devise an offset-based positional training (OPT) strategy, inspired by PoSE [18], that exposes the model to diverse cross-scale combinations without increasing the training sequence length. Theoretically, we show that TPE's hierarchical decomposition provides superior combinatorial coverage of the position space, leads to a tighter generalization bound, and yields more discriminative attention distributions, offering a principled explanation for its extrapolation capability. Our contributions are as follows:

- We propose Tree Rotary Positional Encoding (TPE), which addresses the single-index limitation of rotary encodings by composing bounded, reusable positional components across scales. This tree-structured representation supports long-range positional distinguishability under finite training lengths without semantic segmentation. Controlled HiRoPE-style and no-reset comparisons examine the role of bounded local re-indexing (Appendix D.1).

- We provide a thorough theoretical analysis that formalizes the extrapolation benefits of TPE. We show that TPE achieves superior combinatorial coverage of the positional space and yields attention scores with higher variance across relative positions. These properties explain its stronger generalization and finer positional discriminability in long-context regimes, offering a principled justification for its empirical effectiveness.

- We conduct extensive experiments validating the strong and efficient extrapolation capability of TPE. On a 1.2B model pretrained at length 512, TPE reduces 32K perplexity from 4532.21 to 149.96 under 64× extrapolation. Separately, Passkey accuracy is 58.4% at 8K and 38.1% at 16K. TPE achieves these gains with runtime and memory usage comparable to RoPE.

2 Related Work

Positional encoding is essential for Transformers [19], as it provides sequence order information absent from permutation-invariant attention. Early methods use absolute position embeddings (APEs), including fixed sinusoidal encodings [19] and learned vectors, but they generalize poorly beyond the training length [20]. Relative position embeddings (RPEs) address this issue by modeling pairwise positional relationships [21], typically as attention bias terms, as in T5 [22], ALiBi [12], Kerple [23], and Sandwich [24]. In contrast, rotary position embeddings (RoPE) [16] introduce a distinct approach by rotating query and key vectors such that their inner product inherently encodes relative positional information. Beyond these relative frameworks, recent work explores content-aware position encodings, such as CoPE [25] and DAPE [26], which derive position information from token content. In this line of work, our TPE factorizes positions into a hierarchical multi-scale representation, improving long-range generalization while retaining RoPE's efficiency.

Related to our work, BiPE [27] uses predefined boundaries for intra- and inter-segment positions. HiRoPE [28] was designed specifically for code, using function-level block segmentation, but has also been evaluated on natural-language text with fixed segment splits, making text-based comparisons meaningful. TPE instead composes bounded radix components directly from absolute indices without semantic segmentation. We report matched from-scratch HiRoPE-style baselines and a TPE-no-reset ablation in Appendix D.1.

Long-context modeling faces a key challenge: when LLMs are applied beyond their pre-trained context window, performance often degrades sharply [29] due to out-of-distribution position indices [20]. To address this, Position Interpolation [30] rescales long positions into the pre-trained range, with successors such as NTK-aware scaling, YaRN [14], CLEX [31], Resonance RoPE [32], and LongRoPE [33]. Another class of methods explores position reorganization without additional training by reusing indices within the trained window, including SelfExtend [34], DCA [35], and String [36]. More recently, position simulation enables zero-shot extrapolation by exposing models to long-distance patterns during short-context training, as in RandPos [37], PoSE [18], and later extensions such as Cream [38], LongRecipe [39], and SkipAlign [40]. In contrast, TPE increases effective positional coverage during short-context training via hierarchical decomposition and our OPT strategy, enabling more effective extrapolation to unseen positions. Complementary to positional extrapolation, Dynamic Group Attention (DGA) [41] reduces attention redundancy by aggregating less important tokens while retaining important tokens individually. Core Context Aware (CCA) Attention [42] pools token groups into importance-weighted core representations and combines them with local attention for efficient long-context modeling.

3 Background and Limitations of RoPE

Most modern LLMs [1, 2] encode positional information by applying Rotary Positional Embedding (RoPE) [16], which introduces position-dependent rotations in 2D feature subspaces and enables the incorporation of relative relationships. Given a vector $\mathbf{x}_m \in \mathbb{R}^d$ at position m, RoPE defines

$$f(\mathbf{x}, m) = \mathcal{R}_{\Theta,m}\mathbf{x}, \tag{1}$$

where $\mathcal{R}_{\Theta,m}\in\mathbb{R}^{d\times d}$ is a block-diagonal rotation matrix with angular frequencies $\Theta = (\theta_0, \theta_1, \ldots, \theta_{d/2-1})$, where θ_j rotates the $(2j, 2j+1)$-th 2D subspace. Following the original RoPE formulation, the frequencies are defined as

$$\theta_j = 10000^{-2j/d}, \qquad j \in \{0, 1, \ldots, d/2-1\}. \tag{2}$$

For a position m, the rotation matrix $\mathcal{R}_{\Theta,m}$ is defined as

$$\mathcal{R}_{\Theta,m} = \begin{pmatrix} \cos(m\theta_0) & -\sin(m\theta_0) & \cdots & 0 & 0 \\ \sin(m\theta_0) & \cos(m\theta_0) & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \cos(m\theta_{d/2-1}) & -\sin(m\theta_{d/2-1}) \\ 0 & 0 & \cdots & \sin(m\theta_{d/2-1}) & \cos(m\theta_{d/2-1}) \end{pmatrix} \tag{3}$$

Relative Positional Property of RoPE. The key advantage of RoPE is its ability to naturally encode relative positional information. When computing the self-attention score between a query $\mathbf{q}$ at position m and a key $\mathbf{k}$ at position n, the inner product depends only on their relative distance $m-n$:

$$\langle f(\mathbf{q}, m), f(\mathbf{k}, n)\rangle = (\mathcal{R}_{\Theta,m}\mathbf{q})^\top(\mathcal{R}_{\Theta,n}\mathbf{k}) = \mathbf{q}^\top\mathcal{R}_{\Theta,n-m}\mathbf{k}. \tag{4}$$

This property renders attention translation-invariant, enabling the model to capture the same dependency patterns regardless of their absolute positions.

Limitations of RoPE. 1) Limited extrapolation to long contexts. RoPE encodes relative offsets using periodic terms $\sin((m-n)\theta_j)$ and $\cos((m-n)\theta_j)$. For relative distances far beyond those seen during training, the resulting rotation angles (*i.e.*, phase values) fall outside the training distribution, causing substantial distribution shift [32, 30, 14]. This degrades token ordering behavior and yields inaccurate query–key similarity scores in attention. **2) Under-trained and weakly discriminative low-frequency dimensions.** In RoPE, the rotation angle for each dimension pair $(2j, 2j+1)$ is $m\theta_j$ with $\theta_j = 10000^{-2j/d}$. For large j, θ_j becomes extremely small, so the rotation varies only marginally within the pretraining context. These slow-varying dimensions are therefore insufficiently trained during pretraining and produce highly similar positional representations for nearby relative offsets, limiting fine-grained discrimination. As a result, they contribute little effective positional information, hindering their ability to model long-range dependencies for long sequences [43, 33].

4 Proposed Method

4.1 Motivations and Method Overview

Motivations. The above failure modes of RoPE, *i.e.*, poor long-context extrapolation, under-trained and weakly discriminative low-frequency dimensions, stem from a single structural choice: all positional information is mapped by *one global scalar index*. Concretely, RoPE maps an absolute index m to angles $\{m\theta_j\}_j$, so low-frequency channels (small θ_j) change negligibly during pretraining and remain under-supervised; meanwhile, any attempt to represent positions beyond the training window simply extends the same linear mapping $m \mapsto m\theta_j$ into unseen phase regions, causing a uniform distribution shift of attention scores.

These observations suggest **two explicit design principles**. First, long-range coverage should come from *composing* multiple bounded components whose values are *well covered in the training distribution* (*i.e.*, each component takes many different values during pretraining), instead of extrapolating a single unbounded index. Second, every positional component must vary sufficiently *within* the training window so it receives meaningful supervision. These principles motivate replacing single-index encodings with a structured multi-scale decomposition and adopting training schemes that expose the model to diverse cross-scale combinations.

Method Overview. Motivated by this, we propose **Tree Rotary Positional Encoding (TPE)**, which replaces the single global positional index with a structured multi-scale decomposition. Our TPE represents positions as a composition of bounded-scale components, each encoding only high-frequency variations. This design expresses long-range positions via combinations of components observed during training, rather than by extending a single coordinate beyond its training range. To learn cross-scale interaction patterns in TPE, we further devise an **Offset-based Positional Training (OPT)** strategy. OPT augments training positions with controlled offsets, increasing the diversity of multi-scale positional configurations without requiring explicit long-context pretraining.

4.2 Tree Rotary Positional Encoding

The core idea of TPE is to replace single-scale positional representation with a structured multi-scale one. TPE decomposes an absolute position into multiple bounded components, each associated with its own positional index and operating at a different resolution while encoding only high-frequency positional variations. These bounded components can be combined to represent positional ranges far beyond the training context, while sufficient variation across different hierarchical levels is facilitated by the offset-based positional training strategy introduced below. We put the pseudo-code of TPE in algorithm 1.

Multi-scale Position Decomposition. Standard positional encoding applies a rotation operator $\mathcal{R}_{\Theta,p}$ to a representation vector $\mathbf{x} \in \mathbb{R}^d$ using a single positional index p. In contrast, TPE first replaces this single global index with multiple indices and partitions the feature vector accordingly. Specifically, we split $\mathbf{x}$ into L segments,

$$\mathbf{x} = \left[\mathbf{x}^{(0)}; \mathbf{x}^{(1)}; \ldots; \mathbf{x}^{(L-1)}\right], \qquad \mathbf{x}^{(l)} \in \mathbb{R}^{d/L}, \tag{5}$$

Algorithm 1 Pipeline of Tree Rotary Positional Encoding

Require: Vector $\mathbf{x} \in \mathbb{R}^d$, position p, levels L, radix B, frequency base α (default B), and RoPE frequencies $\Theta = \{\theta_j\}$.

▷ *Step 1: Multi-scale Position Decomposition*

Partition the token $\mathbf{x}$ into L equal-dimensional segments $\mathbf{x} = \big[\mathbf{x}^{(0)}; \mathbf{x}^{(1)}; \ldots; \mathbf{x}^{(L-1)}\big], \quad \mathbf{x}^{(l)} \in \mathbb{R}^{d/L}$

Decompose the absolute position p into L radix-B digits $p^{(l)} = \big\lfloor p/B^l \big\rfloor \bmod B, \quad l = 0, \ldots, L-1$

▷ *Step 2: Scale-weighted Rotary Encoding*

for $l = 0$ to $L-1$ **do**

  Scale frequencies: $\alpha_l \leftarrow \alpha^l$, $\Theta^{(l)} \leftarrow \alpha_l \cdot \Theta_{0:d/(2L)-1}$.

  Apply level-specific rotation: $\mathbf{x}'^{(l)} \leftarrow \mathcal{R}_{\Theta^{(l)}, p^{(l)}}\, \mathbf{x}^{(l)}$.

end for

▷ *Step 3: Multi-scale Positional Encoding Fusion*

Fuse segments: $\mathbf{x}' \leftarrow \mathrm{Concat}\big(\mathbf{x}'^{(0)}, \ldots, \mathbf{x}'^{(L-1)}\big)$

Return: $\mathbf{x}'$

and decompose the absolute position p into L components $\{p^{(0)}, \ldots, p^{(L-1)}\}$ accordingly using a fixed radix B,

$$p^{(l)} = \big\lfloor p/B^l \big\rfloor \bmod B, \qquad l = 0, \ldots, L-1. \tag{6}$$

Here, L can be viewed as the depth of a tree (*i.e.*, hierarchy #levels), B as the number of children at each level, and B^l as its positional resolution. Each $p^{(l)} \in [0, B)$ selects one branch at level l, and the sequence $\{p^{(0)}, \ldots, p^{(L-1)}\}$ specifies a path from the root to a leaf. In this way, each segment $\mathbf{x}^{(l)}$ corresponds to one tree level, and an absolute position is represented by its path through this tree.

Scale-weighted Rotary Encoding. Given segment–index pairs $(\mathbf{x}^{(l)}, p^{(l)})$, a naive approach would apply the transformation $\mathcal{R}_{\Theta, p^{(l)}}\mathbf{x}^{(l)}$ to each segment independently. However, this ignores the hierarchy in which fine scales should vary rapidly while coarse scales evolve more slowly.

To encode this hierarchy, we introduce scale-dependent frequency weighting. Let θ_j denote the base RoPE frequency for the j-th 2D subspace, where $j = 0, \ldots, d/(2L) - 1$ indexes the high-frequency components within each segment, and define for scale l

$$\theta_j^{(l)} = \alpha_l\, \theta_j, \tag{7}$$

which induces a rotation matrix $\mathcal{R}_{\Theta^{(l)}, p^{(l)}}$ with $\Theta^{(l)} = \{\theta_j^{(l)}\}$. The resulting transformation is

$$f^{(l)}(\mathbf{x}^{(l)}, p^{(l)}) = \mathcal{R}_{\Theta^{(l)}, p^{(l)}}\, \mathbf{x}^{(l)}. \tag{8}$$

We use $\alpha_l = \alpha^l$, with $\alpha = B$ by default. Here B controls the radix and reset boundaries; α independently controls frequency scaling. Since $p^{(l)}$ changes only when the absolute position crosses multiples of B^l, larger l capture coarse positional structure, while smaller l encode fine-grained local differences. Combining these scale-weighted rotations yields a multi-scale positional representation that supports both effective training and long-range position coverage.

Multi-scale Positional Encoding Fusion. We form the final representation as:

$$f(\mathbf{x}, p) = \mathrm{Concat}\big(f^{(l)}(\mathbf{x}^{(l)}, p^{(l)}) \;:\; l = 0, \ldots, L-1\big). \tag{9}$$

The hierarchical indices $(p^{(0)}, \ldots, p^{(L-1)})$ span an effective range of size B^L, allowing the fused representation to encode extremely long contexts through coordinated multi-scale contributions rather than extrapolating a single index. This hierarchical combination leads to superior coverage of the positional space and yields a tighter generalization bound than RoPE (see Theorems 1 and 2). Importantly, TPE preserves the same dimensionality and computational cost as standard RoPE.

Unlike RoPE, TPE does not strictly preserve relative-shift invariance. This follows from representing relative displacement as a hierarchical combination of bounded multi-scale components rather than a single global phase difference. While this breaks exact translation invariance, it produces attention scores with higher variance across different relative positions, enhancing *fine-grained positional discrimination* in long contexts (see Theorem 3). Consequently, TPE avoids coupling all rotary dimensions to an extrapolated phase, while the OPT strategy further exposes the model to diverse cross-scale positional configurations during training.

Offset-based Positional Training. While our TPE ensures sufficient positional variation within the training context, learning **cross-scale interaction patterns** remains challenging. Since higher-level

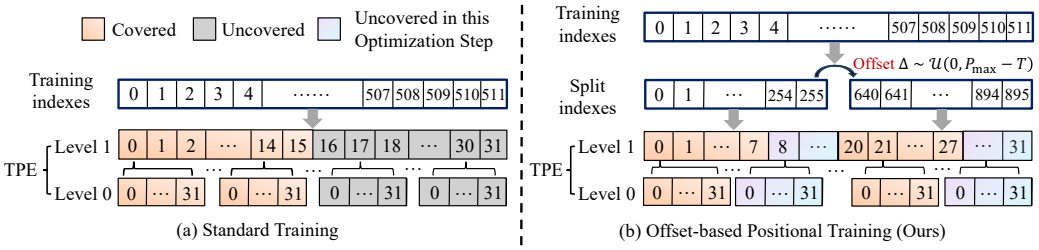


Figure 2: Illustration of offset-based positional training (OPT). (a) In standard training, absolute positions are assigned sequentially, thus many multi-scale index combinations in TPE remain under-trained (uncovered). (b) Our OPT splits each training sequence into two segments and applies a random offset to the second segment. As the offset varies across training samples, the model is exposed to diverse cross-scale positional combinations.For illustration, we use a toy radix $B = 32$.

indices change only when the absolute position advances largely and the standard training strategy exposes the model to only a limited subset of multi-scale index combinations (see Figure 2 (a)). To address this, we devise an offset-based positional training (OPT) strategy inspired by [18] that enriches cross-scale positional supervision without increasing the training sequence length in practice.

In OPT, each training sequence of length T is split into two consecutive chunks of equal length $T/2$. The first chunk is assigned a contiguous index range $[0,\, T/2-1]$, while the second chunk is assigned an offset range $[T/2+\Delta,\, T+\Delta-1]$, where the offset Δ is sampled uniformly from $[0,\, P_{\max}-T]$, with $P_{\max}$ denoting the maximum absolute position considered. The upper bound $P_{\max}-T$ ensures that all assigned indices remain within the valid positional range. Within each chunk, positional indices are contiguous, preserving stable fine-scale positional variation, while the offset between chunks introduces a large positional jump that exposes the model to cross-scale index combinations absent from standard contiguous training (see Figure 2(b)). As Δ is randomly sampled across training samples, the model progressively observes diverse high-level and low-level TPE component combinations, enabling effective learning of cross-scale interactions.

4.3 Theoretical Analysis

TPE employs a hierarchical, multi-scale decomposition of position, fundamentally altering the representation and generalization of positions compared to single-scale encodings RoPE. This yields three key advantages: 1) **superior combinatorial coverage** of the position space, enabling efficient generalization to unseen long contexts; 2) a **tighter generalization bound** due to combinatorial position coverage, and 3) a **more discriminative distribution of attention scores** across relative positions. We formalize these to establish TPE's superiority for long-context modeling.

Theorem 1. ***(Position Coverage)*** *Let $N{=}B^L$ be the position range, T be the block length per training sample, and n be the number of independent samples. Assuming $N{\gg}T$, uniform block offsets, and $L \geq 2$ and $1 \leq T \leq B$, the expected number of distinct absolute positions observed by RoPE after n samples satisfies $C_{\mathrm{RoPE}}(n) \leq N\big(1{-}(1{-}\frac{T}{N})^n\big)$, while TPE's combinatorial representable-encoding coverage satisfies $C_{\mathrm{TPE}}(n) \geq N(1-\frac{1}{B})(1-q_0^n)(1-q^n)^{L-1}$, with $q_0 = 1-\eta T/B,\ q = 1{-}\eta/B,\ \eta = (B{-}1)s/(N{-}T{+}1)$, where $s = N/B$. Under a first-order asymptotic approximation, we have*

$$\frac{C_{\mathrm{TPE}}(n)}{C_{\mathrm{RoPE}}(n)} \gtrsim n^{L-1}\frac{(1-\frac{1}{B})p_0 p^{L-1}}{a}\Big(1-\tfrac{n}{2}(p_0+(L-1)p)+\tfrac{1}{2}na+O(n^2)\Big),$$

where $p_0 = -\log q_0$, $p = -\log q$, and $a = -\log(1-T/N)$.

Theorem 1 quantifies TPE's advantage in covering the position space. For small n, the lower bound on the ratio $\frac{C_{\mathrm{TPE}}(n)}{C_{\mathrm{RoPE}}(n)}$ exhibits n^{L-1} scaling, indicating that TPE can achieve greater combinatorial coverage than RoPE when $L \geq 2$ due to hierarchical decomposition and compositional generalization. For large n, both approach full coverage, while TPE can cover additional positional combinations before saturation, supporting more efficient positional coverage with limited training samples.

Theorem 2. ***(Generalization Bound)*** *Given $\mathcal{P}{=}\{0,\ldots,N{-}1\}$, let $\mathrm{Enc}{:}\mathcal{P}{\to}\mathbb{R}^d$ be the position encoder with $\mathcal{E}{=}\mathrm{Enc}(\mathcal{P})$ of diameter $D{=}\max_{p,p'\in\mathcal{P}}\|\mathrm{Enc}(p){-}\mathrm{Enc}(p')\|_2$. For any h, define*

$\ell(h,p)=\varphi_h(\mathrm{Enc}(p))$ *where* $\varphi_h:\mathbb{R}^d\to[0,1]$ *is* L_ϕ*-Lipschitz. Let* $S \subseteq \mathcal{P}$ *contain* C_{pos} *positions sampled uniformly without replacement. Then,* $\forall\, \delta \in (0,1)$*, with probability at least* $1-\delta$*, we have*

$$\mathbb{E}_{p\sim\mathcal{P}}[\ell(h,p)] - \frac{1}{C_{pos}}\sum_{p\in S}\ell(h,p) \le L_\phi D\sqrt{\frac{\log(1/\delta)}{2C_{pos}}}.$$

Theorem 2 shows that greater observed position coverage yields a tighter generalization bound. Moreover, TPE's hierarchical composition provides greater combinatorial coverage, potentially facilitating generalization to unseen positions.

Theorem 3. ***(Variance of Pre-softmax Attention Scores)*** *Consider* $B\in[256, 2048]$ *and* $L\geq 2$*. Let* $d \geq 512$ *be the embedding dimension and* $M = d/2$*, with* $L \mid M$ *and base* $= 10000$*. The pre-softmax attention scores for a query at* $N-1$ *and a key at* $N-1-\Delta$ *(*$N = B^L$*) under RoPE and TPE are* $S_{\mathrm{RoPE}}(\Delta)=\sum_{r=1}^{M}\left(a_r\cos(\Delta\theta_r)+b_r\sin(\Delta\theta_r)\right)$ *and* $S_{\mathrm{TPE}}(\Delta)=\sum_{j=1}^{L}\sum_{m=1}^{M/L}\Big(a_{j,m}\cos(B^{j-1}\Delta^{(j)}\tilde{\theta}_m)+b_{j,m}\sin(B^{j-1}\Delta^{(j)}\tilde{\theta}_m)\Big)$*, where* a_r*,* b_r*,* $a_{j,m}$*,* $b_{j,m}$ *are query-key coefficients,* $\Delta^{(j)}$ *is the* j*-th base-*B *digit of* Δ*, and* $\tilde{\theta}_m = \theta_{mL} = base^{-mL/M}$*. Assume* Δ *is uniform over* $\{0,\dots,N-1\}$ *and independent of the score coefficients. For* $E \in \{\mathrm{RoPE}, \mathrm{TPE}\}$*, assume zero cross-moments between distinct coefficients, with* $\mathbb{E}[a_{E,r}^2] = \alpha_E > \beta_E = \mathbb{E}[b_{E,r}^2] > 0$*, and* $\alpha_{\mathrm{TPE}}+\beta_{\mathrm{TPE}} > \frac{3L}{2}(\alpha_{\mathrm{RoPE}}+\beta_{\mathrm{RoPE}})$ *(empirically supported). Then the expected variance of the attention score satisfies*

$$\mathbb{E}_{a,b}\,\mathrm{Var}_\Delta\big[S^{\mathrm{TPE}}\big] \;>\; \mathbb{E}_{a,b}\,\mathrm{Var}_\Delta\big[S^{\mathrm{RoPE}}\big].$$

Theorem 3 proves that TPE produces attention scores with higher variance across relative positions than RoPE, yielding more distinctive patterns that help discriminate tokens at different distances. Consistently, Section 5.2 shows TPE-trained models exhibit higher attention weight variance, which mitigates attention dilution in long contexts and enables reliable localization of relevant information.

5 Experiments

Implementation Details.[1] We evaluate TPE by integrating it into decoder-only Transformer models with OLMo [44] at three scales: 60M, 0.3B, and 1.2B. We compare TPE against three representative positional encoding methods: RoPE [16], ALiBi [12], and a strong recent baseline FoPE [17]. All positional encoding methods are pretrained under identical settings. All models are pretrained on a 10B-token subset of the C4-en dataset [22] with a context length of 512 tokens. We use an AdamW optimizer with a peak learning rate of 6.0×10^{-4}, together with a cosine learning-rate schedule and 5,000 warm-up steps. For TPE, we set $L=2$, $B=512$, and $\alpha=512$, giving $\alpha_0=1$ and $\alpha_1=512$, with the maximum OPT sampling range $P_{\max}=256\mathrm{K}$. All pretraining and inference evaluations are conducted on a single node with 16 Ascend 910C accelerators. More experimental details are provided in Appendix B.

5.1 Performance Comparisons on Length Extrapolation

Comparisons on Perplexity In Table 1, we compare perplexity (PPL) under different test length extrapolation. RoPE [16] shows a clear breakdown once extrapolating beyond the training length (*e.g.*, 60M: 35.1→285.5; 0.3B: 24.1→176.9). ALiBi [12] achieves competitive or lower PPL at measured lengths, including 39.8 at 16K for the 60M model versus 80.8 for TPE. Some longer-length entries are unavailable because of OOM in our implementation; we draw no PPL conclusion at those lengths. FoPE [17] substantially slows perplexity degradation, while TPE outperforms RoPE and FoPE at 16K and 32K across the three scales (*e.g.*, 1.2B: 49.9/150.0 versus FoPE's 169.2/610.4). These results characterize likelihood under extrapolation, with retrieval assessed separately. These results indicate that TPE provides more stable long-context extrapolation.

Comparisons on Computational Efficiency. In Table 1, we also report the inference latency and memory at an 8K length. In our implementation, ALiBi uses an explicit attention-score bias and cannot use the same memory-efficient attention path as RoPE/TPE, resulting in higher latency and

[1] Code, checkpoints, and evaluation scripts are available at `https://github.com/AIGCodeOS/aigcode_tpe`.

Table 1: Comparisons with different methods of perplexity under 60M, 0.3B, and 1.2B model. All models are pretrained with 512 context and evaluated from 512 to 32k. `Fwd. Time` and `Mem.` denote inference time (ms) and accelerator memory (GB), measured at length 8192 with batch size 1. – marks an out-of-memory (OOM) result under the same evaluation setting; these entries are not measured PPL values.

#Params	Method	512	1k	2k	4k	8k	16k	32k	Fwd. Time	Mem.
60M	RoPE [16]	35.1	99.1	285.5	506.0	776.0	1056.3	1291.6	32.7	1.9
	ALiBi [12]	35.7	**34.2**	35.8	**38.1**	**39.3**	**39.8**	–	220.0	9.4
	FoPE [17]	**35.1**	34.6	39.9	61.8	128.7	262.2	460.0	36.4	1.9
	TPE (Ours)	35.6	**34.2**	**35.2**	40.9	54.2	80.8	**134.0**	33.4	1.9
0.3B	RoPE [16]	24.1	54.0	176.9	463.8	1113.7	2339.3	3548.8	143.4	3.1
	ALiBi [12]	24.1	**23.0**	24.4	26.2	**27.2**	–	–	924.6	19.3
	FoPE [17]	**24.0**	23.4	25.3	34.1	72.4	202.0	540.6	152.8	3.0
	TPE (Ours)	24.2	23.1	**23.1**	**25.1**	31.2	**47.7**	**108.1**	146.6	3.0
1.2B	RoPE [16]	**20.0**	32.5	107.8	333.7	1101.7	2495.4	4532.2	360.1	6.6
	ALiBi [12]	20.1	19.1	19.8	20.6	**21.0**	–	–	1072.7	22.9
	FoPE [17]	20.0	19.6	20.8	25.8	52.6	169.2	610.4	381.5	6.6
	TPE (Ours)	20.0	**19.0**	**18.9**	**20.0**	24.6	**49.9**	**150.0**	361.1	6.6

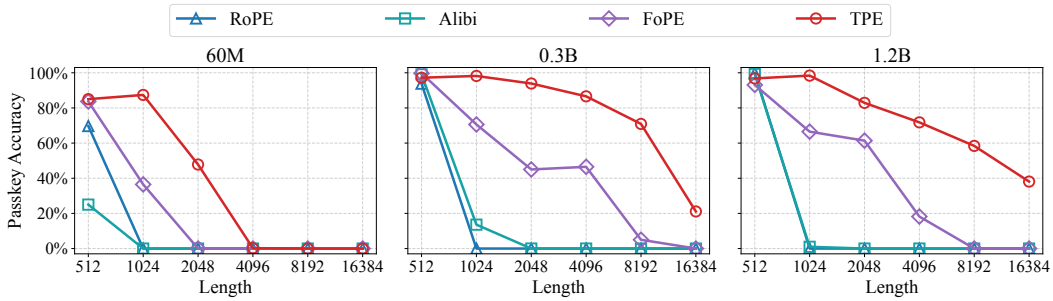


Figure 3: Passkey retrieval accuracy (%) of different methods on 60M, 0.3B, and 1.2B models.

memory usage (60M: 220.0 ms and 9.4 GB versus TPE's 33.4 ms and 1.9 GB). TPE retains runtime and memory usage comparable to RoPE. These implementation-specific costs and OOM entries do not establish ALiBi's PPL at unmeasured lengths.

Comparisons on Passkey Retrieval. In Figure 3, we compare passkey retrieval accuracy under length extrapolation across model scales. RoPE and ALiBi degrade sharply once extrapolating beyond $2\times$ the training length: accuracy drops to nearly zero when context length is greater than 1024. FoPE delays this failure, but its performance decays rapidly with longer contexts. In contrast, TPE exhibits stronger retrieval robustness, with the 0.3B and 1.2B models maintaining high accuracy at $2\times$–$8\times$ extrapolation, while the 60M model degrades earlier. At longer contexts, TPE-1.2B still achieves 58.4%/38.1% accuracy at 8192/16384 tokens.

5.2 Further Experiments

Effect of OPT on different positional encodings. Table 2 examines whether the gains mainly come from OPT. Under the same 0.3B setting, we apply OPT to RoPE and FoPE and compare their passkey retrieval accuracy with TPE. FoPE+OPT performs well at short contexts but drops sharply beyond 4K and collapses at 8K–16K, while RoPE+OPT fails to extrapolate beyond the training length. In contrast, TPE maintains stronger long-context retrieval accuracy, showing that the gains mainly come from the proposed hierarchical positional decomposition rather than OPT.

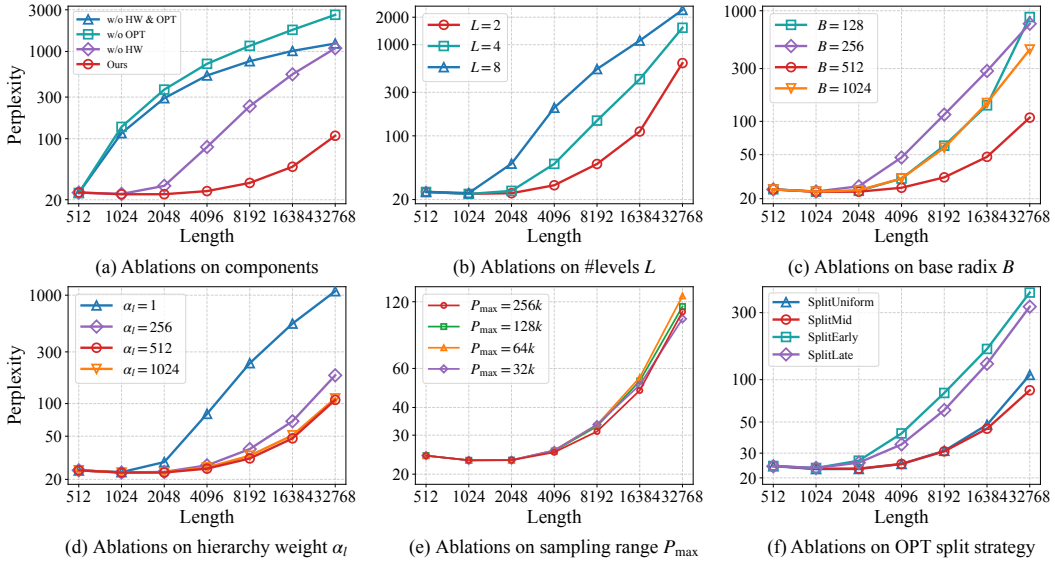
Effect of longer pretraining context. Table 3 evaluates whether TPE remains advantageous when pretrained with a longer 2048-token context. The gap remains clear: RoPE still fails to extrapolate, while FoPE improves at 4K–8K but quickly degrades at 16K and 32K. In contrast, TPE achieves consistently higher accuracy across all lengths and remains strong at 32K, showing robust length extrapolation beyond simply increasing the pretraining context length.

Table 2: Passkey retrieval results on the 0.3B model under different context lengths **with OPT strategy**.

Method	512	1k	2k	4k	8k	16k
RoPE+OPT	58.2	0.0	0.0	0.0	0.0	0.0
FoPE+OPT	98.6	99.1	94.0	11.6	0.0	0.0
TPE (Ours)	97.2	98.2	93.9	86.6	70.9	21.1

Table 3: Passkey retrieval accuracy (%) under **2048-token training** on the 0.3B model.

Method	4k	8k	16k	32k
RoPE	0.0	0.0	0.0	0.0
FoPE	76.2	40.9	0.3	0.0
TPE (Ours)	78.9	72.0	78.9	62.1



(a) Ablations on components (b) Ablations on #levels L (c) Ablations on base radix B

(d) Ablations on hierarchy weight α_l (e) Ablations on sampling range $P_{\max}$ (f) Ablations on OPT split strategy

Figure 4: Ablations of TPE-0.3B (512–32K perplexity) on components, hierarchy depth and radix, frequency scaling, sampling range, and OPT split. HW denotes hierarchy frequency scaling. In (d), the plotted α_l denotes $\alpha_1 = \alpha$, with $\alpha_0 = 1$, $B = 512$, and $L = 2$.

Effect of the Proposed Components. Figure 4(a) analyzes the effects of hierarchy frequency scaling (HW) and Offset Position Training (OPT) on long-context extrapolation. Removing either component makes perplexity increase much faster beyond 2K–4K, indicating that both are important for stable extrapolation. OPT is especially critical at extreme lengths, as removing it causes sharp degradation at 8K–32K. The best and most stable performance is achieved only when HW and OPT are used together, showing that their combination is necessary to fully realize the hierarchical TPE design.

Effect of the number of hierarchy levels L. Figure 4(b) examines the effect of hierarchy depth in TPE. The 2-level configuration achieves the lowest perplexity and degrades slowest as context length increases. Increasing to 4 levels worsens extrapolation beyond 4K, while 8 levels causes rapid collapse and highest far-context perplexity. These results reveal a trade-off between hierarchy depth and extrapolation stability, with shallow hierarchies providing the most robust long-context ability.

Effect of the base radix B. In Figure 4(c), we study the effect of base radix B on long-context perplexity. The setting with B=512 consistently achieves the best performance across all lengths, remaining stable even at 16K–32K tokens. Smaller B (128/256) degrade earlier and exhibit faster perplexity growth, while an overly large setting (B=1024) also underperforms at extreme lengths, suggesting potential over-fragmentation of positional structure.

Effect of hierarchy frequency scaling α. Figure 4(d) varies $\alpha \in \{1, 256, 512, 1024\}$ at fixed $B = 512$ and $L = 2$, with $\alpha_0 = 1$ and $\alpha_1 = \alpha$, independently of the radix and reset boundaries. Setting $\alpha = 1$ causes rapid perplexity growth beyond 4K, whereas $\alpha = 512$ achieves the lowest perplexity from 8K onward. We use $\alpha = B = 512$ by default.

Effect of sampling range $P_{\max}$. In Figure 4(e), we examine the effect of the sampling range on long-context perplexity. Varying the range leads to only modest differences: the curves are nearly indistinguishable up to 8K tokens and diverge slightly as context length increases. At extreme lengths

(16K–32K), the 256K range exhibits the most stable behavior, while shorter ranges (*e.g.*, 32K/64K) trend toward higher perplexity. Thus, we choose sampling range to 256k.

Effect of split strategy in OPT. Figure 4(f) examines different OPT sequence splits: Uniform samples the split from $[1/4,\ 3/4]$ of the sequence length; Mid uses an equal split; Early assigns 20% to the first segment; and Late assigns 80%. Uniform and Mid consistently achieve lower perplexity, indicating that balanced segments better support long-range pattern learning. In contrast, Early and Late splits perform worse at long contexts, suggesting that short or long first segments hinder training.

6 Conclusion

We analyze the limitations of rotary positional encoding stem from its reliance on a single global positional index. To address this, we propose Tree Rotary Positional Encoding (TPE), a structured multi-scale positional representation that enables long-range generalization via compositional reuse of bounded positional components. We further introduce Offset-based Positional Training (OPT) to expose models to diverse cross-scale positional combinations without increasing training length.

Author Contributions

Qiuwu Chen, Ziteng Huang, and Shuhai Zhang contributed equally to this work. Yaofo Chen and Mingkui Tan are the corresponding authors. Qiuwu Chen and Ziteng Huang conceived the main idea of this work. Ziteng Huang, Qiuwu Chen and Shuhai Zhang designed the TPE architecture, developed the theoretical proofs, jointly developed the proposed method, and conducted the experiments. Ziteng Huang, Shuhai Zhang, Zimo Liu, Linxiao Li, Ying Sun, Yuchen Li, and Yifan Zhang contributed to implementation, data preparation, model evaluation, and result analysis. Qiuwu Chen secured the computational resources required for this research. Ziteng Huang and Shuhai Zhang wrote the initial draft. Yaofo Chen and Mingkui Tan supervised the project, provided guidance on methodology and experimental design, and revised the manuscript.

Acknowledgements

This work was supported in part by the Joint Funds of the National Natural Science Foundation of China under Grant No.U24A20327, the Guangdong S&T Program under Grant No.2026B0101110001, the GuanDong Basic and Applied Basic Research Foundation under Grant No.2026A1515010388, the Postdoctoral Fellowship Program of CPSF under Grant GZC20251043. We would also like to thank the Ascend team for their support in providing the computing resources and infrastructure used in this work.

Competing Interests. The authors declare no competing interests.

References

[1] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[2] Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, et al. Qwen2.5-coder technical report. *arXiv preprint arXiv:2409.12186*, 2024.

[3] Jiao Ou, Junda Lu, Che Liu, Yihong Tang, Fuzheng Zhang, Di Zhang, and Kun Gai. Dialogbench: Evaluating llms as human-like dialogue systems. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6137–6170, 2024.

[4] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[5] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*, 2023.

[6] John Yang, Carlos E. Jimenez, Alex L. Zhang, Kilian Lieret, Joyce Yang, Xindi Wu, Ori Press, Niklas Muennighoff, Gabriel Synnaeve, Karthik R. Narasimhan, Diyi Yang, Sida I. Wang, and Ofir Press. Swe-bench multimodal: Do ai systems generalize to visual software domains?, 2024.

[7] Yujia Li, David Choi, Junyoung Chung, Nate Kushman, Julian Schrittwieser, Rémi Leblond, Tom Eccles, James Keeling, Felix Gimeno, Agustin Dal Lago, et al. Competition-level code generation with alphacode. *Science*, 378(6624):1092–1097, 2022.

[8] Raymond Li, Loubna Ben Allal, Yangtian Zi, Niklas Muennighoff, Denis Kocetkov, Chenghao Mou, Marc Marone, Christopher Akiki, Jia Li, Jenny Chim, et al. Starcoder: may the source be with you! *Transactions on Machine Learning Research*, 2023, 2023.

[9] Shengnan An, Zexiong Ma, Zeqi Lin, Nanning Zheng, Jian-Guang Lou, and Weizhu Chen. Make your llm fully utilize the context. *Advances in Neural Information Processing Systems*, 37:62160–62188, 2024.

[10] Zineng Tang, Ziyi Yang, Guoxin Wang, Yuwei Fang, Yang Liu, Chenguang Zhu, Michael Zeng, Cha Zhang, and Mohit Bansal. Unifying vision, text, and layout for universal document processing. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19254–19264. IEEE, 2023.

[11] Dongsheng Wang, Natraj Raman, Mathieu Sibue, Zhiqiang Ma, Petr Babkin, Simerjot Kaur, Yulong Pei, Armineh Nourbakhsh, and Xiaomo Liu. Docllm: A layout-aware generative language model for multimodal document understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8529–8548, 2024.

[12] Ofir Press, Noah A. Smith, and Mike Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. In *International Conference on Learning Representations*, 2022.

[13] Zican Dong, Junyi Li, Xin Men, Xin Zhao, Bingning Wang, Zhen Tian, Ji-Rong Wen, et al. Exploring context window of large language models via decomposed positional vectors. *Advances in Neural Information Processing Systems*, 37:10320–10347, 2024.

[14] Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. Yarn: Efficient context window extension of large language models. In *The Twelfth International Conference on Learning Representations*, 2024.

[15] Shanda Li, Chong You, Guru Guruganesh, Joshua Ainslie, Santiago Ontanon, Manzil Zaheer, Sumit Sanghai, Yiming Yang, Sanjiv Kumar, and Srinadh Bhojanapalli. Functional interpolation for relative positions improves long context transformers. In *The Twelfth International Conference on Learning Representations*, 2024.

[16] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

[17] Ermo Hua, Che Jiang, Xingtai Lv, Kaiyan Zhang, Youbang Sun, Yuchen Fan, Xuekai Zhu, Biqing Qi, Ning Ding, and Bowen Zhou. Fourier position embedding: Enhancing attention's periodic extension for length generalization. In *Forty-second International Conference on Machine Learning*, 2025.

[18] Dawei Zhu, Nan Yang, Liang Wang, Yifan Song, Wenhao Wu, Furu Wei, and Sujian Li. Pose: Efficient context window extension of llms via positional skip-wise training. In *The Twelfth International Conference on Learning Representations*, 2024.

[19] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[20] Amirhossein Kazemnejad, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Payel Das, and Siva Reddy. The impact of positional encoding on length generalization in transformers. *Advances in Neural Information Processing Systems*, 36:24892–24928, 2023.

[21] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 464–468, 2018.

[22] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

[23] Ta-Chung Chi, Ting-Han Fan, Peter Ramadge, and Alexander Rudnicky. Kerple: Kernelized relative positional embedding for length extrapolation. In *Advances in Neural Information Processing Systems*, 2022.

[24] Ta-Chung Chi, Ting-Han Fan, Alexander Rudnicky, and Peter Ramadge. Dissecting transformer length extrapolation via the lens of receptive field analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13522–13537, 2023.

[25] Olga Golovneva, Tianlu Wang, Jason E Weston, and Sainbayar Sukhbaatar. Contextual position encoding: Learning to count what's important. *arXiv preprint arXiv:2405.18719*, 2024.

[26] Chuanyang Zheng, Yihang Gao, Han Shi, Minbin Huang, Jingyao Li, Jing Xiong, Xiaozhe Ren, Michael Ng, Xin Jiang, Zhenguo Li, et al. Dape: Data-adaptive positional encoding for length extrapolation. *Advances in Neural Information Processing Systems*, 37:26659–26700, 2024.

[27] Zhenyu He, Guhao Feng, Shengjie Luo, Kai Yang, Liwei Wang, Jingjing Xu, Zhi Zhang, Hongxia Yang, and Di He. Two stones hit one bird: Bilevel positional encoding for better length extrapolation. In *International Conference on Machine Learning*, pages 17858–17876, 2024.

[28] Kechi Zhang, Ge Li, Huangzhao Zhang, and Zhi Jin. Hirope: Length extrapolation for code models using hierarchical position. In *Annual Meeting of the Association for Computational Linguistics*, pages 13615–13627, 2024.

[29] Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024.

[30] Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. Extending context window of large language models via positional interpolation. *arXiv preprint arXiv:2306.15595*, 2023.

[31] Guanzheng Chen, Xin Li, Zaiqiao Meng, Shangsong Liang, and Lidong Bing. Clex: Continuous length extrapolation for large language models. In *The Twelfth International Conference on Learning Representations*, 2024.

[32] Suyuchen Wang, Ivan Kobyzev, Peng Lu, Mehdi Rezagholizadeh, and Bang Liu. Resonance rope: Improving context length generalization of large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 586–598, 2024.

[33] Yiran Ding, Li Lyna Zhang, Chengruidong Zhang, Yuanyuan Xu, Ning Shang, Jiahang Xu, Fan Yang, and Mao Yang. Longrope: Extending llm context window beyond 2 million tokens. In *Forty-first International Conference on Machine Learning*, 2024.

[34] Hongye Jin, Xiaotian Han, Jingfeng Yang, Zhimeng Jiang, Zirui Liu, Chia-Yuan Chang, Huiyuan Chen, and Xia Hu. Llm maybe longlm: Selfextend llm context window without tuning. In *Proceedings of the 41st International Conference on Machine Learning*, pages 22099–22114, 2024.

[35] Chenxin An, Fei Huang, Jun Zhang, Shansan Gong, Xipeng Qiu, Chang Zhou, and Lingpeng Kong. Training-free long-context scaling of large language models. In *Proceedings of the 41st International Conference on Machine Learning*, pages 1493–1510, 2024.

[36] Chenxin An, Jun Zhang, Ming Zhong, Lei Li, Shansan Gong, Yao Luo, Jingjing Xu, and Lingpeng Kong. Why does the effective context length of llms fall short? In *The Thirteenth International Conference on Learning Representations*, 2025.

[37] Anian Ruoss, Grégoire Delétang, Tim Genewein, Jordi Grau-Moya, Róbert Csordás, Mehdi Bennani, Shane Legg, and Joel Veness. Randomized positional encodings boost length generalization of transformers. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1889–1903, 2023.

[38] Tong Wu, Yanpeng Zhao, and Zilong Zheng. An efficient recipe for long context extension via middle-focused positional encoding. *Advances in Neural Information Processing Systems*, 37:56349–56373, 2024.

[39] Zhiyuan Hu, Yuliang Liu, Jinman Zhao, Suyuchen Wang, Yan Wang, Wei Shen, Qing Gu, Luu Anh Tuan, See Kiong Ng, Zhiwei Jiang, et al. Longrecipe: Recipe for efficient long context generalization in large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11857–11870, 2025.

[40] Wenhao Wu, Yizhong Wang, Yao Fu, Xiang Yue, Dawei Zhu, and Sujian Li. Long context alignment with short instructions and synthesized positions. *arXiv preprint arXiv:2405.03939*, 2024.

[41] Shuhai Zhang, Zeng You, Yaofo Chen, Zhiquan Wen, Qianyue Wang, Zhijie Qiu, Yuanqing Li, and Mingkui Tan. Curse of high dimensionality issue in transformer for long context modeling. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 76600–76624. PMLR, 2025.

[42] Yaofo Chen, Zeng You, Shuhai Zhang, Haokun Li, Yirui Li, Yaowei Wang, and Mingkui Tan. Core context aware transformers for long context language modeling. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 9691–9710. PMLR, 2025.

[43] Xiaoran Liu, Hang Yan, Chenxin An, Xipeng Qiu, and Dahua Lin. Scaling laws of rope-based extrapolation. In *The Twelfth International Conference on Learning Representations*, 2024.

[44] Dirk Groeneveld, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Harsh Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, et al. Olmo: Accelerating the science of language models. *arXiv preprint arXiv:2402.00838*, 2024.

[45] Amirkeivan Mohtashami and Martin Jaggi. Landmark attention: Random-access infinite context length for transformers. *arXiv preprint arXiv:2305.16300*, 2023.

[46] Shanda Li, Chong You, Guru Guruganesh, Joshua Ainslie, Santiago Ontanon, Manzil Zaheer, Sumit Sanghai, Yiming Yang, Sanjiv Kumar, and Srinadh Bhojanapalli. Functional interpolation for relative positions improves long context transformers. In *The Twelfth International Conference on Learning Representations*, 2024.

[47] Jiajun Zhu, Peihao Wang, Ruisi Cai, Jason D Lee, Pan Li, and Zhangyang Wang. Rethinking addressing in language models via contextualized equivariant positional encoding. In *Forty-second International Conference on Machine Learning*, 2025.

APPENDIX

Contents

A Theoretical analysis **15**

A.1 Basic Theoretical results 15

A.2 Proof of Theorem 1 17

A.3 Proof of Theorem 2 21

A.4 Proof of Theorem 3 22

B Implementation Details **29**

C More Discussions of our TPE **30**

C.1 More Comparisons with Bilevel Positional Encoding (BiPE) 30

C.2 More Comparisons with Hierarchical Rotary Position Embedding (HiRoPE). 30

D More Experimental Results **32**

D.1 Controlled Comparisons of Bounded Radix Composition 32

D.2 Native TPE and Post-hoc Adaptation 32

D.3 More Results on 3B Models 33

D.4 A Heterogeneous 20B-token Pretraining Corpus 34

D.5 More Comparisons on RULER 35

D.6 More Comparisons on Downstream Benchmark 35

D.7 Ablation on Offset-based Positional Training 35

D.8 More Comparisons under 2048-token Pretraining 36

D.9 Comparison with Additional Baselines 37

D.10 Checkpoint-level Variability 38

D.11 Performance stability 38

D.12 Visualization of attention weight 39

E Limitations **41**

F Broader Impacts **41**

A Theoretical analysis

Notations. Let B be the base (number of digit values per layer), L the number of TPE layers, $N = B^L$ the total position domain, T the token length sampled per training sample (single-block POSE with block length is T), and n the number of independent training samples. For any fixed digit value at any layer the corresponding absolute-index interval has length $s = N/B$. Assume $N \gg T$ and block start offsets are uniformly random.

For the correspondence with implemented OPT, we reserve T_{train} for the full training sequence and use $t = T_{\text{train}}/2$ for each chunk. The single-block length T in Theorem 1 corresponds to t when considering the random second chunk. The default configuration has $P_{\max} = B^L = N$; the actual two-chunk sampling interval is specified in Remark 1.

A.1 Basic Theoretical results

We start to provide some basic theoretical results, laying the foundation for establishing other theoretical analyses.

Proposition 1. *Let $S_{\text{RoPE}}(\Delta)$ be the pre-softmax attention score for the relative position difference Δ in RoPE, we have*

$$S_{\text{RoPE}}(\Delta) = \sum_{r=1}^{M} [q_{2r-1}k_{2r-1} + q_{2r}k_{2r}] \cos(\omega_r \Delta) + [q_{2r-1}k_{2r} - q_{2r}k_{2r-1}] \sin(\omega_r \Delta).$$

Proof. RoPE transforms query and key vectors through rotation operations in each 2D subspace. Consider the r-th frequency pair corresponding to dimensions $(2r-1, 2r)$. Let $\mathbf{q}_r = q_{2r-1} + iq_{2r}$ and $\mathbf{k}_r = k_{2r-1} + ik_{2r}$ represent the complex view of query and key components in this subspace. The rotary transformation applies position-dependent rotations:

$$\tilde{\mathbf{q}}_r = \mathbf{q}_r e^{im\theta_r}, \quad \tilde{\mathbf{k}}_r = \mathbf{k}_r e^{in\theta_r},$$

where m and n are absolute positions, and θ_r is the frequency. The attention score for $(2r-1, 2r)$ is the real part of the complex inner product:

$$\langle \tilde{\mathbf{q}}_r, \tilde{\mathbf{k}}_r \rangle = \text{Re}(\tilde{\mathbf{q}}_r \overline{\tilde{\mathbf{k}}_r}) = \text{Re}(\mathbf{q}_r \overline{\mathbf{k}_r} e^{i(m-n)\theta_r}). \tag{10}$$

We compute the product $\mathbf{q}_r \overline{\mathbf{k}_r}$ as

$$\mathbf{q}_r \overline{\mathbf{k}_r} = (q_{2r-1} + iq_{2r})(k_{2r-1} - ik_{2r}) = (q_{2r-1}k_{2r-1} + q_{2r}k_{2r}) + i(q_{2r}k_{2r-1} - q_{2r-1}k_{2r}).$$

Let $A_r = q_{2r-1}k_{2r-1} + q_{2r}k_{2r}$, $B_r = q_{2r}k_{2r-1} - q_{2r-1}k_{2r}$, $\Delta = m - n$. Then, we have

$$\mathbf{q}_r \overline{\mathbf{k}_r} e^{i\Delta\theta_r} = (A_r + iB_r)(\cos\Delta\theta_r + i\sin\Delta\theta_r) = (A_r\cos\Delta\theta_r - B_r\sin\Delta\theta_r) + i(A_r\sin\Delta\theta_r + B_r\cos\Delta\theta_r).$$

Taking the real part:

$$\text{Re}[(A_r + iB_r)e^{i\Delta\theta_r}] = A_r\cos\Delta\theta_r - B_r\sin\Delta\theta_r.$$

Substituting back A_r and B_r and combining Eqn. (10), we get

$$\begin{aligned}\langle \tilde{\mathbf{q}}_r, \tilde{\mathbf{k}}_r \rangle &= (q_{2r-1}k_{2r-1} + q_{2r}k_{2r})\cos\Delta\theta_r - (q_{2r}k_{2r-1} - q_{2r-1}k_{2r})\sin\Delta\theta_r \\ &= (q_{2r-1}k_{2r-1} + q_{2r}k_{2r})\cos\Delta\theta_r + (q_{2r-1}k_{2r} - q_{2r}k_{2r-1})\sin\Delta\theta_r.\end{aligned}$$

Summing over all frequency pairs $r = 1$ to M and setting $\omega_r = \theta_r$ yields the final expression:

$$S_{\text{RoPE}}(\Delta) = \sum_{r=1}^{M} \langle \tilde{\mathbf{q}}_r, \tilde{\mathbf{k}}_r \rangle = \sum_{r=1}^{M} [q_{2r-1}k_{2r-1} + q_{2r}k_{2r}] \cos(\omega_r \Delta) + [q_{2r-1}k_{2r} - q_{2r}k_{2r-1}] \sin(\omega_r \Delta).$$

□

Proposition 2. *Let $\ell \geq 1$ be an integer and $\theta \in \mathbb{R}$. Then*

$$\sum_{u=0}^{\ell-1} \cos(2u\theta) = \frac{\sin(\ell\theta)\cos((\ell-1)\theta)}{\sin\theta}, \quad \textit{for } \theta \notin \pi\mathbb{Z}.$$

Consequently,

$$A_\ell(\theta) := \frac{1}{\ell}\sum_{u=0}^{\ell-1} \cos(2u\theta) = \frac{\sin(\ell\theta)\cos((\ell-1)\theta)}{\ell\sin\theta}.$$

Proof. Consider the complex exponential sum:

$$\sum_{u=0}^{\ell-1} e^{i2u\theta} = \frac{1 - e^{i2\ell\theta}}{1 - e^{i2\theta}},$$

where we used the geometric series formula (since $e^{i2\theta} \neq 1$ for $\theta \notin \pi\mathbb{Z}$). Simplify the right-hand side:

$$\frac{1 - e^{i2\ell\theta}}{1 - e^{i2\theta}} = \frac{e^{i\ell\theta}(e^{-i\ell\theta} - e^{i\ell\theta})}{e^{i\theta}(e^{-i\theta} - e^{i\theta})} = \frac{e^{i\ell\theta}(-2i\sin(\ell\theta))}{e^{i\theta}(-2i\sin\theta)} = e^{i(\ell-1)\theta}\frac{\sin(\ell\theta)}{\sin\theta}.$$

Taking the real part on both sides yields

$$\sum_{u=0}^{\ell-1} \cos(2u\theta) = \mathrm{Re}\left(e^{i(\ell-1)\theta}\frac{\sin(\ell\theta)}{\sin\theta}\right) = \frac{\sin(\ell\theta)}{\sin\theta}\cos((\ell-1)\theta).$$

Dividing by ℓ gives the expression for $A_\ell(\theta)$. The formula also holds in the limit when $\theta \in \pi\mathbb{Z}$ by continuity. □

A.2 Proof of Theorem 1

In the following, we first provide two lemmas that will be used for proving Theorem 1.

Lemma 1. *According to the Taylor expansion, when np is small, it holds that*

$$(1-e^{-np})^L \approx (np)^L\Big[1-\tfrac{L}{2}np+\tfrac{L(3L+1)}{24}(np)^2+O((np)^3)\Big].$$

Proof. Let $u=np$. We begin with the Taylor expansion of e^{-u} around $u=0$:

$$e^{-u}=1-u+\frac{u^2}{2}-\frac{u^3}{6}+O(u^4).$$

Thus,

$$1-e^{-u}=u-\frac{u^2}{2}+\frac{u^3}{6}+O(u^4)=u\left(1-\frac{u}{2}+\frac{u^2}{6}+O(u^3)\right).$$

Raising both sides to the power L:

$$(1-e^{-u})^L=u^L\left(1-\frac{u}{2}+\frac{u^2}{6}+O(u^3)\right)^L.$$

Let $w=-\frac{u}{2}+\frac{u^2}{6}+O(u^3)$. Using the binomial expansion:

$$(1+w)^L=1+Lw+\frac{L(L-1)}{2}w^2+O(w^3).$$

Substitute w:

$$w=-\frac{u}{2}+\frac{u^2}{6},\quad w^2=\left(-\frac{u}{2}\right)^2+O(u^3)=\frac{u^2}{4}+O(u^3).$$

Then,

$$(1+w)^L=1+L\left(-\frac{u}{2}+\frac{u^2}{6}\right)+\frac{L(L-1)}{2}\left(\frac{u^2}{4}\right)+O(u^3)=1-\frac{L}{2}u+\left(\frac{L}{6}+\frac{L(L-1)}{8}\right)u^2+O(u^3).$$

Compute the coefficient of u^2:

$$\frac{L}{6}+\frac{L(L-1)}{8}=\frac{4L+3L(L-1)}{24}=\frac{4L+3L^2-3L}{24}=\frac{3L^2+L}{24}=\frac{L(3L+1)}{24}.$$

Therefore,

$$(1-e^{-u})^L=u^L\left[1-\frac{L}{2}u+\frac{L(3L+1)}{24}u^2+O(u^3)\right],$$

which, upon substituting $u=np$, gives the desired result. □

Lemma 2. *Let $R(n)=\frac{(np)^L\left[1-\frac{L}{2}np+\frac{L(3L+1)}{24}(np)^2+O((np)^3)\right]}{na\left[1-\frac{na}{2}+\frac{(na)^2}{6}+O((na)^3)\right]}$, we have*

$$R(n)=n^{L-1}\cdot\frac{p^L}{a}\cdot\left[1-\frac{L}{2}np+\frac{1}{2}na+O(n^2)\right].$$

Proof. We start to factor out $n^{L-1}\frac{p^L}{a}$:

$$R(n)=n^{L-1}\frac{p^L}{a}\cdot\frac{1-\frac{L}{2}np+\frac{L(3L+1)}{24}(np)^2+O((np)^3)}{1-\frac{na}{2}+\frac{(na)^2}{6}+O((na)^3)}.$$

Let $z=na$. The denominator expansion is

$$1-\frac{z}{2}+\frac{z^2}{6}+O(z^3).$$

Its reciprocal can be expanded as

$$\frac{1}{1-\frac{z}{2}+\frac{z^2}{6}+O(z^3)}=1+\frac{z}{2}+\frac{1}{12}z^2+O(z^3).$$

Thus, to first order

$$\frac{1}{1-\frac{na}{2}+\frac{(na)^2}{6}+O((na)^3)} = 1+\frac{na}{2}+O((na)^2).$$

Multiplying the numerator and the reciprocal of the denominator, and keeping terms up to first order in np and na (i.e., ignoring $(np)^2$, $(na)^2$, and $np \cdot na$), we get

$$\begin{aligned}&\left[1-\frac{L}{2}np+\frac{L(3L+1)}{24}(np)^2+O((np)^3)\right]\left[1+\frac{na}{2}+O((na)^2)\right]\\ =&1-\frac{L}{2}np+\frac{na}{2}+O((np)^2,(na)^2,np\cdot na).\end{aligned}$$

Therefore,

$$\begin{aligned}R(n) &= n^{L-1}\frac{p^L}{a}\left[1-\frac{L}{2}np+\frac{1}{2}na+O((np)^2,(na)^2,np\cdot na)\right]\\ &= n^{L-1}\frac{p^L}{a}\left[1-\frac{L}{2}np+\frac{1}{2}na+O(n^2)\right].\end{aligned}$$

□

Theorem 1 ***(Position Coverage)*** *Let $N{=}B^L$ be the position range, T be the block length per training sample, and n be the number of independent samples. Assuming $N{\gg}T$, uniform block offsets, and $L \geq 2$ and $1 \leq T \leq B$, the expected number of distinct absolute positions observed by RoPE after n samples satisfies $C_{\mathrm{RoPE}}(n) \leq N\big(1{-}(1{-}\frac{T}{N})^n\big)$, while TPE's combinatorial representable-encoding coverage satisfies $C_{\mathrm{TPE}}(n) \geq N(1-\frac{1}{B})(1-q_0^n)(1-q^n)^{L-1}$, with $q_0 = 1-\eta T/B,\ q = 1{-}\eta/B,\ \eta = (B{-}1)s/(N{-}T{+}1)$, where $s = N/B$. Under a first-order asymptotic approximation, we have*

$$\frac{C_{\mathrm{TPE}}(n)}{C_{\mathrm{RoPE}}(n)} \gtrsim n^{L-1}\frac{(1-\frac{1}{B})p_0p^{L-1}}{a}\Big(1-\tfrac{n}{2}(p_0+(L-1)p)+\tfrac{1}{2}na+O(n^2)\Big),$$

where $p_0 = -\log q_0$, $p = -\log q$, and $a = -\log(1-T/N)$.

Proof. We prove the theorem mainly using combinatorial probability and asymptotic approximations.

1) We begin to derive the coverage of RoPE. Each block has a starting offset $o \sim \mathrm{Unif}\{0,\dots,N{-}T\}$. For any absolute position $x \in [0, N-1]$, let p_x be its sampling probability under uniform block offsets. Although p_x varies across positions, its average is T/N, since each block contains exactly T positions. After n independent samplings, the probability that x is never sampled is $(1-p_x)^n$. By linearity of expectation and Jensen's inequality, the expected number of distinct positions covered is

$$C_{\mathrm{RoPE}}(n) = \sum_{x=0}^{N-1}\left[1-(1-p_x)^n\right] \leq N\left(1-\left(1-\frac{T}{N}\right)^n\right). \tag{11}$$

2) For TPE coverage, we first focus on TPE per-layer digit coverage. Let $K_\ell(n)$ denote the number of distinct digits observed at level ℓ after n samples. The expected combinatorial coverage is $C_{\mathrm{TPE}}(n) = \mathbb{E}[\prod_{\ell=0}^{L-1} K_\ell(n)]$. To derive a lower bound, consider the valid offset subset $\mathcal{G} = \{0,\dots,H-1\}$, where $H = (B-1)s$, $s = N/B$, and $\eta = H/(N-T+1)$. Since $T \leq B \leq s$, all these offsets are valid. Conditioned on $o \in \mathcal{G}$, its radix-B digits are independent, with B possible values at each lower level and $B-1$ at the highest level.

Let $\underline{K}_\ell(n)$ count the distinct digits from blocks whose starting offsets belong to $\mathcal{G}$. At level 0, all positions within these blocks are considered, while at higher levels, only their starting positions are used. Thus, $\underline{K}_\ell(n) \leq K_\ell(n)$. Since each selected block covers T distinct lowest-level digits, we obtain

$$\begin{aligned}\mathbb{E}[\underline{K}_0(n)] &= B\left(1-\left(1-\frac{\eta T}{B}\right)^n\right),\\ \mathbb{E}[\underline{K}_\ell(n)] &= B\left(1-\left(1-\frac{\eta}{B}\right)^n\right), \quad 1 \leq \ell \leq L-2,\\ \mathbb{E}[\underline{K}_{L-1}(n)] &= (B-1)\left(1-\left(1-\frac{\eta}{B-1}\right)^n\right).\end{aligned}$$

Let $Z \sim \mathrm{Binomial}(n, \eta)$ be the number of selected blocks. Conditioned on Z, the coverage counts $\underline{K}_\ell(n)$ are independent across levels, and their conditional expectations are nondecreasing in Z. Their nonnegative association therefore gives $\mathbb{E}[\prod_\ell \underline{K}_\ell(n)] \geq \prod_\ell \mathbb{E}[\underline{K}_\ell(n)]$. Consequently,

$$C_{\mathrm{TPE}}(n) \geq \prod_{\ell=0}^{L-1} \mathbb{E}[\underline{K}_\ell(n)] \geq N(1-\tfrac{1}{B})(1-q_0^n)(1-q^n)^{L-1}, \tag{12}$$

where $q_0 = 1 - \eta T/B$ and $q = 1 - \eta/B$.

3) We next derive the ratio of the coverages and its approximations. Combining Eqns. (11) and (12), we obtain

$$\begin{aligned} \frac{C_{\mathrm{TPE}}(n)}{C_{\mathrm{RoPE}}(n)} &\geq (1-\tfrac{1}{B})\frac{(1-q_0^n)(1-q^n)^{L-1}}{1-(1-T/N)^n} \\ &= (1-\tfrac{1}{B})\frac{(1-e^{-np_0})(1-e^{-np})^{L-1}}{1-e^{-na}} =: R_{\mathrm{lb}}(n), \end{aligned} \tag{13}$$

where $p_0 = -\log q_0$, $p = -\log q$, and $a = -\log(1 - T/N)$.

For practical intuition, we rewrite the lower bound as $R_{\mathrm{lb}}(n) = (1-\frac{1}{B})\frac{1-e^{-np_0}}{1-e^{-np}}\frac{(1-e^{-np})^L}{1-e^{-na}}$. Using the Taylor expansion, the first factor satisfies

$$\frac{1-e^{-np_0}}{1-e^{-np}} \approx \frac{p_0}{p}\left[1 - \frac{n}{2}(p_0 - p) + O(n^2)\right].$$

For the remaining factor, applying Lemma 1 to the numerator and expanding the denominator as $1 - e^{-na} = na[1 - \frac{na}{2} + O((na)^2)]$, we obtain from Lemma 2

$$\frac{(1-e^{-np})^L}{1-e^{-na}} \approx n^{L-1}\frac{p^L}{a}\left[1 - \frac{L}{2}np + \frac{1}{2}na + O(n^2)\right].$$

Multiplying the two expansions and collecting the first-order terms, we obtain

$$\begin{aligned} R_{\mathrm{lb}}(n) &\approx n^{L-1}\frac{(1-\frac{1}{B})p_0 p^{L-1}}{a}\Big[1 - \frac{n}{2}(p_0 - p) - \frac{L}{2}np + \frac{1}{2}na + O(n^2)\Big] \\ &= n^{L-1}\frac{(1-\frac{1}{B})p_0 p^{L-1}}{a}\Big[1 - \frac{n}{2}(p_0 + (L-1)p) + \frac{1}{2}na + O(n^2)\Big]. \end{aligned}$$

3) We next derive the ratio of the coverages and its approximations. Combining Eqn. (11) and Eqn. (12), we obtain the ratio

$$\frac{C_{\mathrm{TPE}}(n)}{C_{\mathrm{RoPE}}(n)} = \frac{(B(1-q^n))^L}{N\left(1-(1-T/N)^n\right)}.$$

For large N, we can use exponential approximations: $q^n \approx e^{-np}$ and $(1 - T/N)^n \approx e^{-nT/N}$. Then

$$C_{\mathrm{TPE}}(n) \approx N(1-e^{-np})^L, \quad C_{\mathrm{RoPE}}(n) \approx N(1-e^{-nT/N}).$$

Thus, we obtain the ratio

$$\frac{C_{\mathrm{TPE}}(n)}{C_{\mathrm{RoPE}}(n)} \approx \frac{(1-e^{-np})^L}{1-e^{-nT/N}}. \tag{14}$$

For practical intuition, we take the Taylor expansion for the numerator and denominator of the ratio, yielding

$$(1-e^{-np})^L \approx (np)^L\Big[1 - \tfrac{L}{2}np + \tfrac{L(3L+1)}{24}(np)^2 + O((np)^3)\Big]; \tag{15}$$

$$1 - e^{-na} \approx na - \tfrac{(na)^2}{2} + \tfrac{(na)^3}{6} + O((na)^4), \tag{16}$$

where $a = T/N$. The first equation is following Lemma 1. Substituting these into Eqn. (14) and applying Lemma 2, we have

$$\frac{C_{\mathrm{TPE}}(n)}{C_{\mathrm{RoPE}}(n)} \approx n^{L-1}\frac{p^L N}{T}\Big(1 - \tfrac{L}{2}np + \tfrac{1}{2}na + O(n^2)\Big).$$

□

Remark 1 (Correspondence with implemented OPT)**.** *For the default setting $P_{\max} = N$, let $t = T_{\text{train}}/2$. The prefix is $F = \{0, \dots, t-1\}$, and the second chunk is*

$$R_o = \{o, \dots, o+t-1\}, \qquad o = t + \Delta, \quad \Delta \sim \text{Unif}\{0, \dots, N-2t\}.$$

Thus, o is uniform on $\mathcal{O} = \{t, \dots, N-t\}$. This is the random block to which the single-block coverage analysis relates, with block length $T = t$, but its restricted start interval must be distinguished from the idealized interval in the proof.

For level ℓ and digit b, define

$$D_{\ell,b} = \{p \in \{0, \dots, N-1\} : \lfloor p/B^\ell \rfloor \bmod B = b\}.$$

The exact random-chunk hit probability is

$$p_{\ell,b}^{\text{offset}} = \frac{1}{|\mathcal{O}|} \sum_{o \in \mathcal{O}} \mathbf{1}\{R_o \cap D_{\ell,b} \neq \varnothing\}.$$

With $q_{\ell,b}^{\text{prefix}} = \mathbf{1}\{F \cap D_{\ell,b} = \varnothing\}$ and $q_{\ell,b}^{\text{offset}} = 1 - p_{\ell,b}^{\text{offset}}$, the full OPT hit probability is

$$p_{\ell,b}^{\text{OPT}} = 1 - q_{\ell,b}^{\text{prefix}} q_{\ell,b}^{\text{offset}} \geq p_{\ell,b}^{\text{offset}}.$$

The fixed prefix therefore cannot reduce digit coverage relative to the actual random chunk alone. After n independent samples, the expected number of observed digits at level ℓ is

$$\sum_{b=0}^{B-1} \left[1 - \left(q_{\ell,b}^{\text{prefix}} q_{\ell,b}^{\text{offset}}\right)^n\right].$$

The single-block expressions describe the coverage mechanism under their stated sampling assumptions; endpoint and sampling-interval differences should not be treated as exact equalities for implemented OPT. The fixed prefix additionally supplies local and causal supervision.

A.3 Proof of Theorem 2

Next, we present a lemma on concentration under uniform sampling that will be used for Theorem 2.

Lemma 3. *Let $\mathcal{P}$ be a finite set and $f : \mathcal{P} \to \mathbb{R}$ be a function with range $R = \max_{p\in\mathcal{P}} f(p) - \min_{p\in\mathcal{P}} f(p)$. Let $S \subseteq \mathcal{P}$ contain n elements sampled uniformly without replacement. Then, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$,*

$$\mathbb{E}_{p\sim\mathcal{P}}[f(p)] - \frac{1}{n}\sum_{p\in S} f(p) \le R\sqrt{\frac{\log(1/\delta)}{2n}}.$$

Proof. By Hoeffding's inequality for sampling without replacement, the deviation probability satisfies

$$\Pr\left(\mathbb{E}_{p\sim\mathcal{P}}[f(p)] - \frac{1}{n}\sum_{p\in S} f(p) \ge \epsilon\right) \le \exp\left(-\frac{2n\epsilon^2}{R^2}\right).$$

Setting $\epsilon = R\sqrt{\log(1/\delta)/(2n)}$ yields the desired bound. □

Theorem 2 ***(Generalization Bound)*** *Given $\mathcal{P}=\{0,\dots,N-1\}$, let $\mathrm{Enc}:\mathcal{P}\to\mathbb{R}^d$ be the position encoder with $\mathcal{E}=\mathrm{Enc}(\mathcal{P})$ of diameter $D=\max_{p,p'\in\mathcal{P}} \|\mathrm{Enc}(p)-\mathrm{Enc}(p')\|_2$. For any h, define $\ell(h,p)=\varphi_h(\mathrm{Enc}(p))$ where $\varphi_h:\mathbb{R}^d\to[0,1]$ is L_ϕ-Lipschitz. Let $S \subseteq \mathcal{P}$ contain C_{pos} positions sampled uniformly without replacement. Then, $\forall\ \delta \in (0, 1)$, with probability at least $1 - \delta$, we have*

$$\mathbb{E}_{p\sim\mathcal{P}}[\ell(h,p)] - \frac{1}{C_{pos}}\sum_{p\in S} \ell(h,p) \le L_\phi D\sqrt{\frac{\log(1/\delta)}{2C_{pos}}}.$$

Proof. Let $f_h(p) = \ell(h,p)$. By the L_ϕ-Lipschitz property of φ_h, for any $p, p' \in \mathcal{P}$, we have $|f_h(p) - f_h(p')| \le L_\phi\|\mathrm{Enc}(p) - \mathrm{Enc}(p')\|_2 \le L_\phi D$. Thus, the range of f_h is bounded by $R_h \le L_\phi D$.

Applying Lemma 3 with $f = f_h$, $R = R_h$, and $n = C_{pos}$, we obtain, with probability at least $1 - \delta$,

$$\mathbb{E}_{p\sim\mathcal{P}}[\ell(h,p)] - \frac{1}{C_{pos}}\sum_{p\in S} \ell(h,p) \le R_h\sqrt{\frac{\log(1/\delta)}{2C_{pos}}} \le L_\phi D\sqrt{\frac{\log(1/\delta)}{2C_{pos}}}.$$

This completes the proof. □

A.4 Proof of Theorem 3

We present two important lemmas before proving Theorem 3.

Lemma 4. *Let the attention score be given by $S(\Delta) = \sum_{i=1}^{n} (a_i \cos(\Delta\theta_i) + b_i \sin(\Delta\theta_i))$, where the pairs (a_i, b_i) are random variables satisfying $\mathbb{E}[a_i a_j] = \alpha\delta_{ij}$, $\mathbb{E}[b_i b_j] = \beta\delta_{ij}$, and $\mathbb{E}[a_i b_j] = 0$ for all i, j, with $\alpha, \beta > 0$ and δ_{ij} denoting the Kronecker delta. Then*

$$\mathbb{E}[S(\Delta)^2] = \sum_{i=1}^{n} \left(\alpha \cos^2(\Delta\theta_i) + \beta \sin^2(\Delta\theta_i)\right).$$

Moreover, if Δ is sampled independently of the coefficients, the expected positional variance satisfies

$$\mathbb{E}_{a,b}[\mathrm{Var}_\Delta(S)] = \sum_{i=1}^{n} \left(\alpha \,\mathrm{Var}_\Delta[\cos(\Delta\theta_i)] + \beta \,\mathrm{Var}_\Delta[\sin(\Delta\theta_i)]\right).$$

Both identities also apply to TPE by replacing $\Delta\theta_i$ with its corresponding digit-based rotary phase.

Proof. Expanding the square $S(\Delta)^2$:

$$\begin{aligned} S(\Delta)^2 = \sum_{i,j} (&a_i a_j \cos(\Delta\theta_i)\cos(\Delta\theta_j) + a_i b_j \cos(\Delta\theta_i)\sin(\Delta\theta_j) \\ &+ b_i a_j \sin(\Delta\theta_i)\cos(\Delta\theta_j) + b_i b_j \sin(\Delta\theta_i)\sin(\Delta\theta_j)). \end{aligned}$$

Taking expectation and using the given conditions, all cross-terms with $i \neq j$ vanish by the zero cross-moment assumptions. For $i = j$, we have

$$\mathbb{E}[a_i^2]\cos^2(\Delta\theta_i) + \mathbb{E}[b_i^2]\sin^2(\Delta\theta_i) + 2\mathbb{E}[a_i b_i]\cos(\Delta\theta_i)\sin(\Delta\theta_i).$$

Since $\mathbb{E}[a_i b_i] = 0$ and $\mathbb{E}[a_i^2] = \alpha$, $\mathbb{E}[b_i^2] = \beta$, we get the result:

$$\mathbb{E}[S(\Delta)^2] = \sum_{i=1}^{n} \left(\alpha \cos^2(\Delta\theta_i) + \beta \sin^2(\Delta\theta_i)\right).$$

Next, by the definition of variance,

$$\mathrm{Var}_\Delta(S) = \mathbb{E}_\Delta[S(\Delta)^2] - (\mathbb{E}_\Delta[S(\Delta)])^2.$$

Taking the expectation of $S(\Delta)$ over Δ, we obtain

$$\mathbb{E}_\Delta[S(\Delta)] = \sum_{i=1}^{n} \left(a_i \mathbb{E}_\Delta[\cos(\Delta\theta_i)] + b_i \mathbb{E}_\Delta[\sin(\Delta\theta_i)]\right).$$

Squaring this expression and taking the expectation over a, b, the zero cross-moment assumptions eliminate all cross-terms, yielding

$$\mathbb{E}_{a,b}\left[(\mathbb{E}_\Delta[S(\Delta)])^2\right] = \sum_{i=1}^{n} \left(\alpha(\mathbb{E}_\Delta[\cos(\Delta\theta_i)])^2 + \beta(\mathbb{E}_\Delta[\sin(\Delta\theta_i)])^2\right).$$

Subtracting this term from $\mathbb{E}_\Delta\mathbb{E}_{a,b}[S(\Delta)^2]$, we obtain

$$\begin{aligned} \mathbb{E}_{a,b}[\mathrm{Var}_\Delta(S)] &= \sum_{i=1}^{n} \Big(\alpha \left(\mathbb{E}_\Delta[\cos^2(\Delta\theta_i)] - (\mathbb{E}_\Delta[\cos(\Delta\theta_i)])^2\right) \\ &\qquad + \beta \left(\mathbb{E}_\Delta[\sin^2(\Delta\theta_i)] - (\mathbb{E}_\Delta[\sin(\Delta\theta_i)])^2\right)\Big) \\ &= \sum_{i=1}^{n} \left(\alpha \,\mathrm{Var}_\Delta[\cos(\Delta\theta_i)] + \beta \,\mathrm{Var}_\Delta[\sin(\Delta\theta_i)]\right). \end{aligned}$$

The derivation also applies to TPE by replacing $\Delta\theta_i$ with its corresponding digit-based rotary phase. □

Lemma 5. *For any integer $\ell \geq 1$ and $\theta \in (0, 1]$, we have*

$$|A_\ell(\theta)| \leq \min\left(1, \frac{2}{\ell\theta}\right),$$

where $A_\ell(\theta) = \frac{1}{\ell}\sum_{u=0}^{\ell-1}\cos(2u\theta)$.

Proof. We prove the two bounds separately.

1) For the bound $|A_\ell(\theta)| \leq 1$, since $|\cos(2u\theta)| \leq 1$ for all u, the triangle inequality gives

$$|A_\ell(\theta)| = \left|\frac{1}{\ell}\sum_{u=0}^{\ell-1}\cos(2u\theta)\right| \leq \frac{1}{\ell}\sum_{u=0}^{\ell-1}|\cos(2u\theta)| \leq \frac{1}{\ell}\cdot\ell = 1.$$

2) For the bound $|A_\ell(\theta)| \leq \frac{2}{\ell\theta}$, using the closed-form expression for the sum of cosines in Proposition 2, we have

$$A_\ell(\theta) = \frac{\sin(\ell\theta)\cos((\ell-1)\theta)}{\ell\sin\theta}.$$

Taking absolute values and using $|\cos((\ell-1)\theta)| \leq 1$, we obtain

$$|A_\ell(\theta)| \leq \frac{|\sin(\ell\theta)|}{\ell|\sin\theta|} \leq \frac{1}{\ell|\sin\theta|}.$$

For $\theta \in (0, 1] \subset (0, \pi/2]$, the inequality $\sin\theta \geq \theta/2$ holds. Indeed, consider the function $f(\theta) = \sin\theta - \theta/2$ on $[0, \pi/2]$. We have $f(0) = 0$ and $f'(\theta) = \cos\theta - 1/2$. For $\theta \in [0, \pi/3]$, $f'(\theta) \geq 0$, so f is non-decreasing; for $\theta \in [\pi/3, \pi/2]$, $\cos\theta \leq 1/2$, but we can check the endpoint: $f(\pi/2) = 1 - \pi/4 > 0$. Thus $f(\theta) \geq 0$ on $[0, \pi/2]$. Hence,

$$\frac{1}{|\sin\theta|} \leq \frac{2}{\theta}.$$

Substituting into the bound for $|A_\ell(\theta)|$ yields

$$|A_\ell(\theta)| \leq \frac{2}{\ell\theta}.$$

Combining the two bounds, we conclude

$$|A_\ell(\theta)| \leq \min\left(1, \frac{2}{\ell\theta}\right).$$

□

Theorem 3 ***(Variance of Pre-softmax Attention Scores)*** *Consider $B \in [256, 2048]$ and $L \geq 2$. Let $d \geq 512$ be the embedding dimension and $M = d/2$, with $L \mid M$ and base $= 10000$. The pre-softmax attention scores for a query at $N-1$ and a key at $N-1-\Delta$ ($N = B^L$) under RoPE and TPE are*

$$S_{\mathrm{RoPE}}(\Delta) = \sum_{r=1}^{M} \left(a_r \cos(\Delta\theta_r) + b_r \sin(\Delta\theta_r)\right),$$

$$S_{\mathrm{TPE}}(\Delta) = \sum_{j=1}^{L} \sum_{m=1}^{M/L} \left(a_{j,m} \cos(B^{j-1}\Delta^{(j)}\tilde{\theta}_m) + b_{j,m} \sin(B^{j-1}\Delta^{(j)}\tilde{\theta}_m)\right),$$

where a_r, b_r, $a_{j,m}$, $b_{j,m}$ are query-key coefficients, $\Delta^{(j)}$ is the j-th base-B digit of Δ, and $\tilde{\theta}_m = \theta_{mL} = \text{base}^{-mL/M}$. Assume Δ is uniform over $\{0, \ldots, N-1\}$ and independent of the score coefficients. For $E \in \{\mathrm{RoPE}, \mathrm{TPE}\}$, assume zero cross-moments between distinct coefficients, with $\mathbb{E}[a_{E,r}^2] = \alpha_E > \beta_E = \mathbb{E}[b_{E,r}^2] > 0$, and $\alpha_{\mathrm{TPE}} + \beta_{\mathrm{TPE}} > \frac{3L}{2}(\alpha_{\mathrm{RoPE}} + \beta_{\mathrm{RoPE}})$ (empirically supported). Then the expected variance of the attention score satisfies

$$\mathbb{E}_{a,b}\mathrm{Var}_\Delta[S^{\mathrm{TPE}}] > \mathbb{E}_{a,b}\mathrm{Var}_\Delta[S^{\mathrm{RoPE}}].$$

Proof. Since $N-1$ has digit $B-1$ at every level, the position pair $(N-1, N-1-\Delta)$ satisfies $p_q^{(j-1)} - p_k^{(j-1)} = \Delta^{(j)}$, yielding the stated TPE attention score.

1) We first derive the expected positional variances for RoPE and TPE. Define

$$\mu_K(\omega) = \frac{1}{K}\sum_{u=0}^{K-1} e^{iu\omega}, \qquad A_K(\omega) = \frac{1}{K}\sum_{u=0}^{K-1} \cos(2u\omega).$$

Using the definition of variance and the trigonometric identities, we obtain

$$U_K(\omega) = \mathrm{Var}_u[\cos(u\omega)] = \mathbb{E}_u[\cos^2(u\omega)] - (\mathbb{E}_u[\cos(u\omega)])^2 = \frac{1 + A_K(\omega)}{2} - (\mathrm{Re}\,\mu_K(\omega))^2,$$

$$W_K(\omega) = \mathrm{Var}_u[\sin(u\omega)] = \mathbb{E}_u[\sin^2(u\omega)] - (\mathbb{E}_u[\sin(u\omega)])^2 = \frac{1 - A_K(\omega)}{2} - (\mathrm{Im}\,\mu_K(\omega))^2,$$

where u is uniformly distributed over $\{0, \ldots, K-1\}$. Adding the two expressions and using $\cos^2(u\omega) + \sin^2(u\omega) = 1$ gives

$$D_K(\omega) := U_K(\omega) + W_K(\omega) = 1 - |\mu_K(\omega)|^2. \tag{17}$$

This quantity captures the combined positional variation of the cosine and sine components and will be used to bound the TPE variance.

By Lemma 4 and the zero cross-moment assumptions, the expected positional variance for RoPE is

$$V_{\mathrm{RoPE}} := \mathbb{E}_{a,b}[\mathrm{Var}_\Delta(S_{\mathrm{RoPE}})] = \sum_{r=1}^{M} \left(\alpha_{\mathrm{RoPE}} U_N(\theta_r) + \beta_{\mathrm{RoPE}} W_N(\theta_r)\right).$$

Similarly, since Δ is uniform over $\{0, \ldots, B^L - 1\}$, each digit $\Delta^{(j)}$ is uniform over $\{0, \ldots, B-1\}$. Applying the same lemma to the TPE phases yields

$$V_{\mathrm{TPE}} := \mathbb{E}_{a,b}[\mathrm{Var}_\Delta(S_{\mathrm{TPE}})] = \sum_{j=1}^{L} \sum_{m=1}^{M/L} \left(\alpha_{\mathrm{TPE}} U_B(B^{j-1}\tilde{\theta}_m) + \beta_{\mathrm{TPE}} W_B(B^{j-1}\tilde{\theta}_m)\right).$$

2) We derive a lower bound for the TPE variance. We first show that for any $B \geq 256$ and $\theta \in [10^{-4}, 1]$, at least one frequency $\omega \in \{\theta, B\theta\}$ satisfies

$$U_B(\omega) > \frac{7}{20}, \qquad D_B(\omega) > \frac{7}{8}.$$

Let $x = B\theta$. We consider two cases.

Case 1: $x \geq 6$. We take $\omega = \theta$. By the closed-form expressions for the trigonometric sums,

$$|\mu_B(\theta)| = \frac{|\sin(x/2)|}{B\sin(\theta/2)}, \qquad |A_B(\theta)| \leq \frac{|\sin x|}{B\sin\theta}.$$

For $\theta \in (0, 1]$, the inequalities $\sin(\theta/2) \geq 23\theta/48$ and $\sin\theta \geq \theta \sin 1$ imply

$$|\mu_B(\theta)| \leq \frac{48|\sin(x/2)|}{23x} \leq \frac{48}{23x} \leq \frac{8}{23} < \frac{1}{\sqrt{8}},$$

and

$$|A_B(\theta)| \leq \frac{|\sin x|}{x \sin 1}.$$

Consequently,

$$\begin{aligned} D_B(\theta) &= 1 - |\mu_B(\theta)|^2 > \frac{7}{8}, \\ U_B(\theta) &= \frac{1 + A_B(\theta)}{2} - (\operatorname{Re}\mu_B(\theta))^2 \\ &\geq \frac{1}{2} - \frac{|\sin x|}{2x\sin 1} - \left(\frac{48|\sin(x/2)|}{23x}\right)^2. \end{aligned}$$

To bound the last expression, we divide $x \geq 6$ into three intervals. For $x \in [6, 7]$, we have $|\sin x| \leq 0.66$ and $|\sin(x/2)| \leq 0.36$; for $x \in [7, 8]$, we have $|\sin x| \leq 1$ and $|\sin(x/2)| \leq 0.76$; and for $x \geq 8$, both sine terms are bounded by 1. Substituting these bounds into the preceding inequality yields

$$\begin{aligned} U_B(\theta) &\geq \frac{1}{2} - \frac{0.66}{12\sin 1} - \left(\frac{48 \cdot 0.36}{23 \cdot 6}\right)^2 > 0.418, \quad x \in [6, 7], \\ U_B(\theta) &\geq \frac{1}{2} - \frac{1}{14\sin 1} - \left(\frac{48 \cdot 0.76}{23 \cdot 7}\right)^2 > 0.363, \quad x \in [7, 8], \\ U_B(\theta) &\geq \frac{1}{2} - \frac{1}{16\sin 1} - \left(\frac{48}{23 \cdot 8}\right)^2 > 0.357, \quad x \geq 8. \end{aligned}$$

Therefore, $U_B(\theta) > 7/20$ holds for all $x \geq 6$.

Case 2: $x < 6$. We take $\omega = x = B\theta$. If $x \leq 1$, then $Bx = B^2\theta \geq 256^2 \times 10^{-4} > 6$. Since $x \in (0, 1]$, the bounds established in Case 1 apply to $\omega = x$.

For $1 < x < 6$, the closed-form expression for $\mu_B(x)$ gives

$$|\mu_B(x)|^2 = \frac{\sin^2(Bx/2)}{B^2\sin^2(x/2)} \leq \frac{1}{256^2\sin^2 3} < 0.0008,$$

where we used $\sin(x/2) \geq \sin 3$ for $x \in [1, 6]$. To bound $A_B(x)$, let $\delta \in [0, \pi/2]$ be the distance from x to the nearest multiple of π. By periodicity and symmetry, $A_B(x) = A_B(\delta)$.

If $\delta \leq 4/B$, let $f(t) = \cos(2B\delta t)$. Since $|f'(t)| \leq 2B\delta$, the difference between the left Riemann sum and the corresponding integral satisfies

$$\begin{aligned} \left|A_B(\delta) - \int_0^1 f(t)\,dt\right| &= \left|\sum_{u=0}^{B-1} \int_{u/B}^{(u+1)/B} (f(u/B) - f(t))\,dt\right| \\ &\leq \sum_{u=0}^{B-1} \int_{u/B}^{(u+1)/B} 2B\delta\left(t - \frac{u}{B}\right)dt = \delta. \end{aligned}$$

Moreover, $\sin y/y \geq -1/4$ for $y > 0$. Indeed, the ratio is nonnegative for $y \in (0, \pi]$; for $y \in [\pi, 4]$, using $\sin y \geq -(y - \pi)$ gives $\sin y/y \geq \pi/4 - 1 > -1/4$; and for $y \geq 4$, we have $\sin y/y \geq -1/y \geq -1/4$. Therefore,

$$\begin{aligned} A_B(\delta) &\geq \int_0^1 \cos(2B\delta t)\,dt - \delta \\ &= \frac{\sin(2B\delta)}{2B\delta} - \delta \\ &\geq -\frac{1}{4} - \frac{4}{B} > -0.266, \end{aligned}$$

where the ratio at $\delta = 0$ is defined by continuity.

Otherwise, if $\delta > 4/B$, the closed-form expression for A_B and the monotonicity of $\sin\delta$ on $[0, \pi/2]$ give

$$A_B(\delta) \geq -\frac{1}{B\sin\delta} \geq -\frac{1}{B\sin(4/B)} > -0.251.$$

Combining both cases, we have $A_B(x) > -0.28$. Therefore,

$$\begin{aligned} U_B(x) &= \frac{1 + A_B(x)}{2} - (\operatorname{Re}\mu_B(x))^2 \\ &> \frac{1 - 0.28}{2} - 0.0008 > \frac{7}{20}, \\ D_B(x) &= 1 - |\mu_B(x)|^2 > 1 - 0.0008 > \frac{7}{8}. \end{aligned}$$

Hence, for every $\theta \in [10^{-4}, 1]$, at least one of θ and $B\theta$ satisfies both desired bounds.

Now, each TPE base frequency $\tilde{\theta}_m = 10000^{-mL/M}$ lies in $[10^{-4}, 1]$. For each m, select one of the first two levels whose frequency $\omega_m \in \{\tilde{\theta}_m, B\tilde{\theta}_m\}$ satisfies the bounds above. Since all other positional variance terms are nonnegative, Eqn. (17) gives

$$\begin{aligned} V_{\mathrm{TPE}} &\geq \sum_{m=1}^{M/L} (\alpha_{\mathrm{TPE}} U_B(\omega_m) + \beta_{\mathrm{TPE}} W_B(\omega_m)) \\ &= \sum_{m=1}^{M/L} ((\alpha_{\mathrm{TPE}} - \beta_{\mathrm{TPE}}) U_B(\omega_m) + \beta_{\mathrm{TPE}} D_B(\omega_m)) \\ &> \frac{M}{L}\left(\frac{7}{20}(\alpha_{\mathrm{TPE}} - \beta_{\mathrm{TPE}}) + \frac{7}{8}\beta_{\mathrm{TPE}}\right) \\ &\geq \frac{7M}{20L}(\alpha_{\mathrm{TPE}} + \beta_{\mathrm{TPE}}). \end{aligned} \tag{18}$$

3) We next derive an upper bound for the RoPE variance. Since the positional variance is no greater than the corresponding second moment, Lemma 4 gives

$$\begin{aligned} V_{\mathrm{RoPE}} &\leq \sum_{r=1}^{M} \left(\alpha_{\mathrm{RoPE}} \mathbb{E}_\Delta[\cos^2(\Delta\theta_r)] + \beta_{\mathrm{RoPE}} \mathbb{E}_\Delta[\sin^2(\Delta\theta_r)]\right) \\ &= \frac{M(\alpha_{\mathrm{RoPE}} + \beta_{\mathrm{RoPE}})}{2} + \frac{\alpha_{\mathrm{RoPE}} - \beta_{\mathrm{RoPE}}}{2} \sum_{r=1}^{M} A_N(\theta_r). \end{aligned}$$

By Lemma 5, $|A_N(\theta_r)| \leq 2/(N\theta_r)$, since $\theta_r \in (0, 1]$. Using $\theta_r = 10000^{-r/M}$, we obtain

$$\frac{1}{M}\sum_{r=1}^{M} |A_N(\theta_r)| \leq \frac{2}{N} \cdot \frac{1}{M} \sum_{r=1}^{M} 10000^{r/M}.$$

Since 10000^x is increasing on $[0, 1]$, its right Riemann sum is bounded by the integral plus the endpoint difference divided by M. Thus,

$$\frac{1}{M}\sum_{r=1}^{M} 10000^{r/M} \leq \int_0^1 10000^x \, dx + \frac{10000 - 1}{M} = \frac{9999}{\ln 10000} + \frac{9999}{M} < 1200,$$

where $M \geq 256$. Since $N = B^L \geq 256^2$, it follows that

$$\frac{1}{M}\sum_{r=1}^{M} |A_N(\theta_r)| < \frac{2400}{N} \leq \frac{2400}{256^2} < \frac{1}{20}.$$

Substituting this bound into the variance expression and using $\alpha_{\mathrm{RoPE}} > \beta_{\mathrm{RoPE}} > 0$, we obtain

$$\begin{aligned} V_{\mathrm{RoPE}} &< \frac{M}{2}(\alpha_{\mathrm{RoPE}} + \beta_{\mathrm{RoPE}}) + \frac{M}{40}(\alpha_{\mathrm{RoPE}} - \beta_{\mathrm{RoPE}}) \\ &\leq \frac{21M}{40}(\alpha_{\mathrm{RoPE}} + \beta_{\mathrm{RoPE}}). \end{aligned} \tag{19}$$

Table 4: Empirical validation of $\mathbb{E}[a_r^2] > \mathbb{E}[b_r^2]$ over 100 samples for RoPE and TPE on a 0.3B model.

Position	$\mathbb{E}[a_r^2]$	$\mathbb{E}[b_r^2]$	$\mathbb{E}[a_r^2] - \mathbb{E}[b_r^2]$
RoPE			
2000	3.34	1.61	1.73
4000	3.62	1.73	1.90
8000	4.14	2.08	2.06
TPE			
2000	19.38	12.60	6.78
4000	40.49	20.92	19.58
8000	92.28	39.29	52.99

Finally, under the assumed condition $\alpha_{\mathrm{TPE}} + \beta_{\mathrm{TPE}} > \frac{3L}{2}(\alpha_{\mathrm{RoPE}} + \beta_{\mathrm{RoPE}})$, combining Eqs. (18) and (19) gives

$$V_{\mathrm{TPE}} > \frac{7M}{20L}(\alpha_{\mathrm{TPE}} + \beta_{\mathrm{TPE}}) > \frac{21M}{40}(\alpha_{\mathrm{RoPE}} + \beta_{\mathrm{RoPE}}) > V_{\mathrm{RoPE}}.$$

□

Empirical evidence for the second-moment assumptions in Theorem 3. To verify the second-moment assumptions in Theorem 3, we measure $\mathbb{E}[a_r^2]$ and $\mathbb{E}[b_r^2]$ over 100 samples at positions 2000, 4000, and 8000. As shown in Table 4, for both RoPE and our TPE, $\mathbb{E}[a_r^2]$ consistently exceeds $\mathbb{E}[b_r^2]$ across all tested positions, with the gap widening at longer distances, supporting the within-method inequality. Moreover, the ratios of the summed second moments between TPE and RoPE are 6.46, 11.48, and 21.15, respectively, exceeding the required factor of 3 for $L = 2$ and empirically supporting the additional condition in Theorem 3. A possible reason for this bias lies in the structure of a_r and b_r: $a_r = q_{2r-1}k_{2r-1} + q_{2r}k_{2r}$ sums in-phase products of matched channels, while $b_r = q_{2r}k_{2r-1} - q_{2r-1}k_{2r}$ is an anti-symmetric cross-term. When paired query-key components are positively correlated, the sum in a_r reinforces signal energy, whereas the subtraction in b_r often leads to cancellation, resulting in a smaller second moment.

Visualization of Attention Weights between RoPE and our TPE. Figure 5 shows the mean cumulative attention weight assigned to tokens beyond the 1k-th position in the final layer of a 0.3B model on 2k-length sequences. A clear contrast is observed: RoPE's cumulative weights quickly saturate near 1.0 soon after position 1k, indicating that much of the attention mass is assigned to earlier positions in the displayed range. In contrast, TPE's weights increase gradually and rise sharply only near the 2048-th position, indicating a stronger emphasis on positions closer to the current query. The variance of attention weights in the final layer is about 20 times larger for TPE (4.77×10^{-5}) than for RoPE (2.37×10^{-6}), demonstrating more differentiated attention allocation across positions. This observation is qualitatively consistent with the positional discrimination studied in Theorem 3, although the theorem concerns pre-softmax attention scores. Thus, TPE exhibits more position-selective attention patterns in long contexts, supporting its ability to distinguish tokens at different distances.

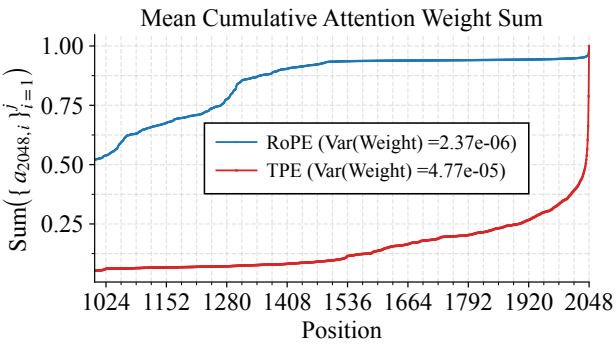


Figure 5: Comparison of cumulative attention weights between TPE and RoPE. We plot the cumulative sum of attention weights assigned to tokens beyond the 1k-th position in the final layer on a 0.3B model with 2k sequences. The legend reports the attention-weight variance in the last layer.

B Implementation Details

Hardware. All pretraining runs and inference-time evaluations are conducted on a single node equipped with 16 Ascend 910C accelerators, each with 64 GB device memory.

Framework. We adopt the open-source OLMo [44] pretraining framework. OLMo employs a decoder-only Transformer architecture and incorporates architectural and training refinements—including RMSNorm, rotary positional embeddings (RoPE), and Z-loss regularization—to improve convergence stability and hardware utilization in large-scale pretraining. We follow the official distributed training setup provided by OLMo for model pretraining.

Data. We use a subset of the C4-en [22] pretraining corpus officially supported by OLMo. C4-en is constructed from Common Crawl web snapshots and processed via language identification, extensive cleaning/filtering, and deduplication to obtain a large-scale English web-text corpus. The resulting training subset contains 10.65B tokens. We evaluate on the pre-defined held-out split of C4-en, totaling 1.02M tokens.

Pretraining setup. We pretrain OLMo-style models at four scales (60M, 0.3B, 1.2B and 3B parameters) with a maximum sequence length of 512. We use AdamW with a peak learning rate of 6.0×10^{-4} and a cosine learning-rate schedule with 5,000 warm-up steps. To keep the total optimization budget consistent across settings, all runs are trained for 20,307 steps with a global batch size of 1,024. Detailed architectural hyperparameters are reported in Table 5.

Table 5: Model configurations for different scales.

Model Scale	60M	0.3B	1.2B	3B
Num Heads	8	16	16	24
Num Layers	8	16	16	28
MLP Ratio	8	8	8	8
Head Dim	64	64	128	128
Mini Batchsize	64	32	16	8

TPE configuration. Unless otherwise specified, TPE uses $L = 2$, $B = 512$, and frequency weights $\alpha_l = \alpha^l$ with $\alpha = B = 512$; hence $\alpha_0 = 1$ and $\alpha_1 = 512$. OPT is enabled by default with $P_{\max} = 256\text{K}$.

Hyperparameter selection. We recommend using the training context and intended position range as the primary guide: start with $B \approx T_{\text{train}}$, choose the smallest L such that $B^L \geq P_{\max}$, and use $\alpha = B$ by default. The offset range should cover the intended evaluation or deployment length. For $T_{\text{train}} = 512$ and $P_{\max} = 256\text{K}$, $B = 512$ and $L = 2$ cover the range with two active levels; larger depths divide the rotary dimensions among additional, effectively inactive levels. The same defaults are effective across the evaluated 60M–3B models. For a new context setting, a small sweep around $B \approx T_{\text{train}}$, $\alpha \approx B$, and the minimum feasible depth is a practical starting point.

Evaluation. We consider two kinds of evaluation metrics: 1) We compute perplexity on the C4-en validation split using the final training checkpoint. We evaluate sequence lengths from 512 to 32,768; for each length ℓ, we set the validation chunk size to ℓ, feed the resulting chunks to the model, and compute perplexity via standard autoregressive likelihood. 2) We evaluate length-wise retrieval using a passkey task [45] where the key is a numeric string drawn from 1–5000. The key is randomly inserted into a distractor paragraph at five relative positions (0%, 20%, 40%, 60%, and 80%). Given the resulting prompt, the model generates up to 10 tokens, and we extract the predicted key for exact-match scoring. For each insertion position we evaluate 200 samples; thus, for each sequence length we test 1,000 samples in total and report the mean accuracy (optionally also per-position accuracies) as the retrieval performance at that length.

C More Discussions of our TPE

C.1 More Comparisons with Bilevel Positional Encoding (BiPE)

BiPE [27] decomposes positional information into intra-segment and inter-segment components: it injects an additive absolute encoding inside each segment and applies a multiplicative relative encoding. We compare TPE with BiPE from both methodological and empirical perspectives.

From a methodological standpoint, the two methods differ in two important ways.

- First, BiPE mixes additive absolute and multiplicative relative encodings, which can lead to less coherent interactions between the two levels. In BiPE, the intra-segment position is added to the input embeddings (additive APE), while the inter-segment position acts on attention scores (multiplicative RPE). TPE, in contrast, represents all positional components with a unified multi-scale rotary encoding—a purely relative scheme applied consistently across scales. This yields a homogeneous parameterization that is inherently more favorable for long-range length extrapolation.
- Second, BiPE depends on predefined segment boundaries (e.g., periods, newlines) and its performance is sensitive to segmentation quality. TPE operates directly on the absolute token index and requires no sentence, paragraph, or other semantic boundary information, removing the extra assumption that reliable segmentation is available and making it more generally applicable.

Empirically, we train both models from scratch under the same setting with a context length of 512 and evaluate their length extrapolation. As shown in Tables 25 and 26, TPE extrapolates much more effectively to long contexts in terms of both PPL and retrieval accuracy, confirming the practical advantage of the multi-scale rotary design over the mixed additive-multiplicative scheme of BiPE.

C.2 More Comparisons with Hierarchical Rotary Position Embedding (HiRoPE).

HiRoPE [28] extends rotary position embedding into a hierarchical format tailored for source code. It augments RoPE with function-level and token-level indices extracted from the syntactic structure of code, aiming to improve length extrapolation for code models without extra training. We clarify the relationship between TPE and HiRoPE from both design motivation and generality.

From a method perspective, TPE is more general than HiRoPE. HiRoPE is motivated by the hierarchical nature of source code and adapts RoPE to a specific, code-aware structure. In contrast, TPE starts from a broader diagnosis: standard rotary encodings are fundamentally limited because all positional dimensions are governed by a *single* global index, which tightly couples training-time positional variation with inference-time extrapolation. TPE addresses this single-index limitation directly, without assuming any domain-specific segmentation. Here, generality refers to constructing positions directly from absolute indices without semantic segmentation; HiRoPE's applicability to natural-language text is discussed below.

Furthermore, although HiRoPE introduces a hierarchical code-aware index, its token-level branch still inherits RoPE's index out-of-distribution (OOD) problem. In HiRoPE, the low-level token index is not reset when the function index changes; for instance, the combined representation follows the pattern $((0,0),(0,1),\ldots,(0,4),(1,5),(1,6))$, meaning the token-level index keeps increasing unboundedly, just like the original RoPE index. Consequently, this branch remains exposed to RoPE-style long-range OOD behavior. By contrast, TPE resolves this issue through a bounded multi-scale decomposition with local reset: the local component is bounded and resets within each group, yielding representations such as $((0,0),(0,1),\ldots,(0,4),(1,0),(1,1))$. No single local index grows unboundedly across groups; long positions are represented by recombining local components rather than extrapolating an ever-increasing token index. For clarity, these illustrative pairs list the coarse component first and the token or local component second. The reset example uses a toy group size of five; the experimental TPE configuration uses $B = 512$.

Finally, HiRoPE is designed specifically for code data and relies on function-level code block segmentation to construct its positional hierarchy. This describes its code-oriented formulation. HiRoPE introduces hierarchical rotary dimension allocation and has also been evaluated on natural-

language long-text modeling with fixed segment splits. The relevant distinction is therefore the positional parameterization rather than whether a comparison on text is possible.

We evaluate HiRoPE-style variants under matched from-scratch OLMo-300M training with 512-token context and approximately 10B tokens. They split rotary dimensions equally between the unbounded token index p and the segment index $\lfloor p/s \rfloor$, with $s \in \{64, 128\}$. TPE instead uses bounded radix components and is further compared with a no-reset variant. Tables 6 and 7 show that hierarchical splitting alone does not reproduce TPE's long-range retrieval gains. These are controlled HiRoPE-style baselines rather than a reproduction of every setting in the original HiRoPE study. The results support bounded local re-indexing within TPE, without establishing modulo reset as uniquely optimal among alternative bounded mappings.

D More Experimental Results

D.1 Controlled Comparisons of Bounded Radix Composition

We train matched OLMo-300M models from scratch with a 512-token context and approximately 10B tokens. The HiRoPE-style variants split rotary dimensions equally between p and $\lfloor p/s \rfloor$, with $s = 64$ or 128, and apply the hierarchy to all query–key pairs. TPE-no-reset retains TPE's feature partitioning, scale weighting, and OPT, but uses $(p, \lfloor p/512 \rfloor)$ instead of resetting the finest-level index to $p \bmod 512$.

Tables 6 and 7 separate likelihood from retrieval. At 32K, TPE-no-reset has lower PPL than TPE (29.3 versus 108.1), but its Passkey accuracy is zero from 2K onward. The HiRoPE-style variants also lose retrieval at 8K–16K. These controls show that hierarchical splitting and OPT alone do not reproduce TPE's retrieval gains and support bounded local re-indexing within the evaluated construction. They do not establish modulo as superior to all alternative bounded mappings.

Table 6: Perplexity of matched HiRoPE-style baselines and the no-reset ablation on OLMo-300M pretrained at length 512.

Method	512	1K	2K	4K	8K	16K	32K
TPE (Ours)	24.2	23.1	23.1	25.1	31.2	47.7	108.1
HiRoPE-style ($s = 64$)	24.1	25.0	30.3	50.1	115.8	275.3	711.6
HiRoPE-style ($s = 128$)	24.1	25.0	30.3	49.8	113.1	271.2	632.7
TPE-no-reset	24.5	23.5	23.3	23.6	24.3	25.8	29.3

Table 7: Passkey retrieval accuracy (%) for the matched OLMo-300M comparisons.

Method	512	1K	2K	4K	8K	16K
TPE (Ours)	97.2	98.2	93.9	86.6	70.9	21.1
HiRoPE-style ($s = 64$)	93.9	53.3	70.8	54.3	0.0	0.0
HiRoPE-style ($s = 128$)	98.7	80.7	61.9	52.8	0.0	0.0
TPE-no-reset	76.0	19.5	0.0	0.0	0.0	0.0

D.2 Native TPE and Post-hoc Adaptation

For the post-hoc comparison, RoPE+YaRN and RoPE→TPE start from the same RoPE-pretrained OLMo-1.2B checkpoint and extend continued-training context from 512 to 1K. RoPE+YaRN uses $2\times$ YaRN, whereas RoPE→TPE replaces RoPE with $L = 2$, $B = 512$, $\alpha = 512$, and OPT. Tables 8 and 9 also report native RoPE, native TPE, and TPE+YaRN.

RoPE→TPE substantially improves likelihood under extrapolation, reaching 29.6 PPL at 32K, compared with 150.0 for native TPE. However, native TPE retains stronger retrieval at 8K/16K (58.4%/38.1% versus 19.8%/2.6%). TPE+YaRN further improves the reported TPE results, reaching 73.7 PPL at 32K and 62.5% Passkey accuracy at 16K. Thus, adaptation provides a useful migration path, while native pretraining has a retrieval advantage under the evaluated adaptation budget. This comparison does not prove that from-scratch training is necessary or that a longer post-hoc training schedule cannot close the gap.

Table 8: Perplexity of native and continued-training variants on OLMo-1.2B.

Method	512	1K	2K	4K	8K	16K	32K
RoPE	20.0	32.5	107.8	333.7	1101.7	2495.4	4532.2
TPE	20.0	19.0	18.9	20.0	24.6	49.9	150.0
RoPE+YaRN	20.8	19.6	36.2	127.5	402.8	1174.1	2975.5
TPE+YaRN	20.9	19.7	19.2	19.5	21.7	32.7	73.7
RoPE→TPE	21.1	20.1	19.7	20.0	20.6	22.3	29.6

Table 9: Passkey retrieval accuracy (%) of native and continued-training OLMo-1.2B variants.

Method	512	1K	2K	4K	8K	16K
RoPE	99.9	0.0	0.0	0.0	0.0	0.0
TPE	96.8	98.4	82.9	71.8	58.4	38.1
RoPE+YaRN	99.6	97.4	0.0	0.0	0.0	0.0
TPE+YaRN	96.2	99.0	96.4	95.1	89.5	62.5
RoPE→TPE	99.4	99.5	85.3	56.2	19.8	2.6

D.3 More Results on 3B Models

In Tables 10–13, we further evaluate whether the advantage of TPE can be preserved when scaling the model size to 3B parameters. As shown in Table 10, RoPE suffers from severe perplexity degradation once the evaluation length exceeds the 512-token training context, and its perplexity increases dramatically at 16K and 32K tokens. In contrast, TPE maintains stable perplexity across all evaluated lengths, achieving 24.8 at 8K, 41.9 at 16K, and 94.7 at 32K. This indicates that the hierarchical positional decomposition remains effective under larger model scaling and provides substantially more stable long-context extrapolation than standard RoPE.

Table 10: PPL comparison on the 3B model.

Model	512	1k	2k	4k	8k	16k	32k
RoPE	17.7	27.0	76.3	169.9	472.8	1025.2	13281.6
TPE (ours)	18.3	17.5	17.7	19.6	24.8	41.9	94.7

Table 11 further reports passkey retrieval accuracy on the 3B model. RoPE collapses immediately beyond the training length, dropping from 84.2 at 512 tokens to 0.0 at 1K and longer contexts. By comparison, TPE achieves strong retrieval performance at short and medium extrapolation lengths, maintaining 97.6 at 1K, 85.8 at 2K, and 58.8 at 4K. Although the retrieval accuracy also decreases at very long contexts, TPE still demonstrates a clear advantage over RoPE, showing that the proposed positional representation better supports long-range information access after scaling to 3B parameters.

Table 11: Passkey retrieval accuracy (%) on the 3B model.

Model	512	1k	2k	4k	8k	16k
RoPE	84.2	0.0	0.0	0.0	0.0	0.0
TPE (ours)	99.7	97.6	85.8	58.8	2.7	0.0

In Table 12, we additionally examine short-context downstream task performance. TPE remains broadly comparable to RoPE across common downstream benchmarks, with better results on Basic, OBQA, SciQ, and CSQA, while RoPE performs slightly better on several other tasks such as Wino, Hella, PIQA, ARC-easy, and NQ. These results suggest that the improved extrapolation ability of TPE does not come at the cost of a substantial degradation in general short-context capability.

Table 12: Performance on short-context downstream tasks on the 3B model.

Model	Basic	SIQA	Wino	OBQA	SciQ	Hella	PIQA	CSQA	ARC-easy	NQ
RoPE	8.5	34.9	51.1	25.8	25.1	26.4	52.2	19.8	24.9	6.5
TPE (ours)	8.7	33.8	50.4	27.0	25.9	23.7	50.6	19.9	23.7	5.4

Finally, Table 13 presents RULER results on the 3B model. TPE generally outperforms RoPE on retrieval-oriented subtasks under extrapolated context lengths. In particular, TPE achieves 69.8 and

63.5 on `niah_s` at 1K and 2K, respectively, while RoPE drops from 28.1 at 1K to 0.0 at 2K. TPE also maintains non-zero performance on multiple subtasks where RoPE fails, including `niah_mq`, `niah_mv`, `qa_1`, and `vt`. At 4K, where the benchmark becomes much more challenging, TPE still achieves non-zero results on several subtasks, whereas RoPE collapses across all tasks. Overall, these results suggest that TPE provides stronger long-context robustness than RoPE when scaling to the 3B model size.

Table 13: Comparisons on RULER. "–" denotes that the subtask corpus cannot be constructed at the corresponding context length.

Length	Method	niah_s	niah_mk	niah_mq	niah_mv	cwe	fwe	qa_1	qa_2	vt
1k	RoPE	28.1	4.2	0.0	12.5	0.0	0.0	3.1	–	0.0
1k	TPE (ours)	69.8	9.4	21.9	18.8	0.0	68.8	18.8	–	6.3
2k	RoPE	0.0	0.0	0.0	0.0	0.0	0.0	0.0	–	0.0
2k	TPE (ours)	63.5	10.4	21.9	40.6	0.0	0.0	12.5	–	18.8
4k	RoPE	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4k	TPE (ours)	12.5	1.0	0.0	6.3	3.1	0.0	12.5	15.6	0.0

D.4 A Heterogeneous 20B-token Pretraining Corpus

To examine data diversity and training budget beyond the original C4-en setting, we train 60M models on a 20.05B-token mixture of C4-en, OLMo-MixBooks, and OLMo-MixWiki. Table 14 gives the effective token composition.

Tables 15 and 16 show improved extrapolation for TPE on this mixture. At 32K, TPE achieves 135.8 PPL, versus 518.7 for FoPE and 1652.5 for RoPE. TPE retains 14.9% Passkey accuracy at 8K and 1.1% at 16K, while the other methods score zero at those lengths. These are relative improvements at the 60M scale; the 16K result does not establish high absolute retrieval performance.

Table 14: Effective token composition of the heterogeneous 20.05B-token corpus used for the 60M experiment.

Datasource	Effective tokens	Proportion
C4-en	9.85B	49.13%
OLMo-MixBooks	5.26B	26.21%
OLMo-MixWiki	4.94B	24.66%
Total	20.05B	100%

Table 15: Perplexity of 60M models pretrained on the 20.05B-token mixed corpus.

Method	512	1K	2K	4K	8K	16K	32K
RoPE	36.4	110.3	345.1	641.4	968.9	1291.5	1652.5
FoPE	36.4	35.9	42.0	68.1	144.4	292.3	518.7
TPE (Ours)	36.9	35.3	37.0	45.1	60.7	87.4	135.8

Table 16: Passkey retrieval accuracy (%) of 60M models pretrained on the 20.05B-token mixed corpus.

Method	512	1K	2K	4K	8K	16K
RoPE	63.4	0.0	0.0	0.0	0.0	0.0
FoPE	69.1	29.8	0.0	0.0	0.0	0.0
TPE (Ours)	92.3	89.4	71.2	32.8	14.9	1.1

D.5 More Comparisons on RULER

In Table 17, we report additional RULER results on the 1.2B model. TPE consistently shows stronger robustness under length extrapolation, especially on retrieval-oriented NIAH tasks. At 1K and 2K contexts, TPE substantially outperforms RoPE and FoPE on `niah_s`, maintaining 100.0 and 96.9 accuracy, while RoPE rapidly collapses and FoPE degrades more noticeably. At 4K, although all methods become more challenging, TPE still achieves the best results on most subtasks, indicating that the proposed hierarchical positional representation better preserves long-context retrieval ability under difficult multi-task evaluation.

Table 17: More results on the RULER benchmark with different test length. "–" denotes that the subtask corpus cannot be constructed at the corresponding context length.

Length	Method	niah_s	niah_mk	niah_mq	niah_mv	cwe	fwe	qa_1	qa_2	vt
512	RoPE	100.0	17.7	40.6	15.6	15.6	0.0	21.9	–	–
512	FoPE	97.9	32.3	34.3	31.2	21.9	0.0	21.9	–	–
512	TPE(ours)	100.0	26.0	37.5	31.2	43.8	0.0	21.9	–	–
1K	RoPE	18.7	1.0	3.1	40.6	0.0	31.2	0.0	–	0.0
1K	FoPE	63.6	2.1	3.1	15.6	0.0	15.6	3.1	–	0.0
1K	TPE(ours)	100.0	15.6	34.4	25.0	0.0	62.5	18.8	–	0.0
2K	RoPE	0.0	0.0	0.0	0.0	0.0	0.0	0.0	–	0.0
2K	FoPE	64.6	6.2	9.4	18.8	0.0	0.0	3.1	–	3.1
2K	TPE(ours)	96.9	13.5	21.9	25.0	0.0	6.2	9.4	–	3.1
4K	RoPE	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4K	FoPE	3.1	0.0	0.0	0.0	0.0	0.0	0.0	6.2	0.0
4K	TPE(ours)	28.1	4.2	3.1	9.4	3.1	0.0	3.1	9.4	0.0

D.6 More Comparisons on Downstream Benchmark

In Table 18, we evaluate the downstream performance of the 1.2B base pretrained models. Since these models are not instruction-finetuned, the absolute scores are naturally limited, and our focus is whether TPE preserves general task performance while improving length extrapolation. In the updated 13-task evaluation, TPE achieves the best scores on SIQA, PIQA, ARC-E, M-STEM, and M-Hum., and ties FoPE on SciQ.

The average scores are 39.8 for RoPE, 39.1 for FoPE, and 39.3 for TPE. These results suggest that TPE largely preserves general downstream ability while providing substantially stronger long-context extrapolation.

Table 18: Updated short-context benchmark results (%) on OLMo-1.2B. Avg. is the reported mean across the 13 preceding task columns. Best scores in each column are bold.

Method	**BA**	**SIQA**	**Wino**	**OBQA**	**SciQ**	**Hella**	**PIQA**	**CSQA**	**ARC-E**	**NQ**	**M-STEM**	**M-Social**	**M-Hum.**	**Avg.**
RoPE	**8.4**	45.8	51.7	**39.6**	61.9	45.6	69.8	**51.4**	55.7	4.3	24.7	**31.3**	27.8	**39.8**
FoPE	4.9	46.7	**54.0**	37.2	**62.4**	**45.7**	69.0	46.7	56.2	**4.5**	24.4	29.6	27.4	39.1
TPE (Ours)	6.2	**46.9**	52.8	36.4	**62.4**	45.3	**70.2**	45.6	**58.1**	3.9	**25.7**	29.6	**28.0**	39.3

D.7 Ablation on Offset-based Positional Training

In Tables 19 and 20, we analyze whether the improvement mainly comes from the OPT strategy. Applying OPT to RoPE does not improve extrapolation and even leads to worse perplexity at long contexts, while RoPE+OPT still fails on passkey retrieval beyond the training length. FoPE+OPT improves short-range retrieval, but its performance drops sharply at 4K and collapses at 8K–16K tokens. In contrast, TPE maintains the lowest perplexity and the strongest passkey accuracy across long contexts, showing that OPT alone does not account for the long-context gains.

RoPE+OPT increases long-context PPL and reduces 512-token Passkey accuracy from 93.8% to 58.2%. FoPE+OPT improves retrieval at 1K–2K but falls to 11.6% at 4K and 0.0% at 8K–16K. TPE reaches 86.6%, 70.9%, and 21.1% at 4K, 8K, and 16K. OPT therefore interacts differently with the positional parameterization: it exposes combinations of bounded components in TPE, whereas applying offsets to a single-index rotary encoding does not reproduce the same long-range behavior. Together with the matched HiRoPE-style and no-reset comparisons in Appendix D.1, these results support bounded radix composition with OPT.

Table 19: PPL results on the 0.3B model under different context lengths.

Method	512	1k	2k	4k	8k	16k	32k
RoPE	24.1	54.0	176.9	463.8	1113.7	2339.3	3548.8
RoPE+OPT	24.5	101.7	392.8	963.9	2133.8	3781.9	5008.7
FoPE	24.0	23.3	25.3	34.1	72.4	202.0	540.6
FoPE+OPT	24.1	22.9	23.6	36.2	99.5	326.8	1046.1
TPE (Ours)	24.2	23.1	23.1	25.1	31.2	47.7	108.1

Table 20: Passkey retrieval results on the 0.3B model under different context lengths.

Method	512	1k	2k	4k	8k	16k
RoPE	93.8	0.0	0.0	0.0	0.0	0.0
RoPE+OPT	58.2	0.0	0.0	0.0	0.0	0.0
FoPE	99.7	70.6	45.0	46.5	5.0	0.0
FoPE+OPT	98.6	99.1	94.0	11.6	0.0	0.0
TPE (Ours)	97.2	98.2	93.9	86.6	70.9	21.1

D.7.1 Fixed Prefix versus Random Origin in OPT

We compare standard OPT with TPE-shift, which randomly offsets both chunks while preserving their gap. Standard OPT retains a fixed low-index prefix and supervises offset queries attending to earlier prefix keys under causal masking.

Tables 21 and 22 show that randomizing the origin does not improve the evaluated OLMo-300M setting. At 32K, PPL increases from 108.1 to 399.1; at 4K, Passkey accuracy decreases from 86.6% to 17.7%, and TPE-shift reaches 0.0% at 8K–16K. These results support the fixed-prefix construction under the tested conditions.

Table 21: Perplexity of standard OPT and the random-origin TPE-shift variant on OLMo-300M.

Method	512	1K	2K	4K	8K	16K	32K
TPE (standard)	24.2	23.1	23.1	25.1	31.2	47.7	108.1
TPE-shift	24.3	23.2	24.5	38.0	77.0	170.2	399.1

Table 22: Passkey retrieval accuracy (%) for the OPT origin ablation on OLMo-300M.

Method	512	1K	2K	4K	8K	16K
TPE (standard)	97.2	98.2	93.9	86.6	70.9	21.1
TPE-shift	94.4	81.7	63.1	17.7	0.0	0.0

D.8 More Comparisons under 2048-token Pretraining

In Tables 23 and 24, we further evaluate models pretrained with a longer 2048-token context. Under this stronger setting, TPE achieves perplexity comparable to FoPE across 4K–32K tokens, while

RoPE still suffers from severe degradation. More importantly, TPE shows a clear advantage in passkey retrieval: it maintains high accuracy from 4K to 32K, whereas FoPE rapidly drops after 8K and fails at 32K. These results show that TPE remains highly effective for length extrapolation even when the pretraining context is increased.

Table 23: PPL under 2048-token training.

Model	4k	8k	16k	32k
RoPE	63.6	257.7	682.7	1592.8
FoPE	21.9	23.2	29.8	64.1
TPE (Ours)	21.8	22.9	32.6	69.2

Table 24: Passkey retrieval accuracy (%) under 2048-token training.

Model	4k	8k	16k	32k
RoPE	0.0	0.0	0.0	0.0
FoPE	76.2	40.9	0.3	0.0
TPE (Ours)	78.9	72.0	78.9	62.1

D.9 Comparison with Additional Baselines

Perplexity comparison with more baselines. In Table 25, we compare TPE with additional positional encoding baselines on the 0.3B model. Several methods perform reasonably near the training length but degrade rapidly as the context becomes longer. YaRN and BiPE improve over vanilla RoPE, but their perplexity still grows substantially at 8K–32K tokens. FoPE is the strongest baseline among these methods, yet TPE consistently achieves lower perplexity across all extrapolated lengths, especially at 16K and 32K. This demonstrates the stronger stability of TPE under extreme length extrapolation.

Table 25: Perplexity comparison with additional baselines on the 0.3B model. "–" indicates GPU memory overflow under the same evaluation setting.

Method	512	1k	2k	4k	8k	16k	32k
RoPE [16]	24.1	54.0	176.9	463.8	1113.7	2339.3	3548.8
YaRN [14]	24.9	23.5	24.1	33.8	123.0	779.4	2726.8
NoPE [20]	25.0	86.7	617.2	2085.2	4796.7	8988.2	14945.8
FiRE [46]	24.3	25.6	179.1	1056.5	3306.2	–	–
TAPE [47]	24.0	57.0	176.9	396.1	865.3	1946.8	–
BiPE [27]	24.5	23.4	28.3	97.2	344.9	949.3	1864.4
FoPE [17]	24.0	23.3	25.3	34.1	72.4	202.0	540.6
TPE (Ours)	24.2	23.1	23.1	25.1	31.2	47.7	108.1

Passkey retrieval comparison with more baselines. In Table 26, we report passkey retrieval accuracy with additional baselines on the 0.3B model. Most methods fail to maintain retrieval accuracy once the context exceeds 2K or 4K tokens, including FiRE, TAPE, BiPE, and RoPE+YaRN. FoPE delays this collapse but still drops to 5.0 at 8K and 0.0 at 16K. In contrast, TPE preserves high retrieval accuracy across all evaluated lengths and remains non-trivial even at 16K tokens, further confirming its superior robustness for long-context information retrieval.

Table 26: Passkey retrieval accuracy comparison with additional baselines on the 0.3B model. TPE remains substantially more robust at long contexts than the added baselines.

Method	512	1k	2k	4k	8k	16k
RoPE [16]	93.8	0.0	0.0	0.0	0.0	0.0
YaRN [14]	93.5	64.5	0.4	0.0	0.0	0.0
NoPE [20]	42.8	0.2	0.0	0.0	0.0	0.0
FiRE [46]	98.3	93.8	78.6	0.0	0.0	0.0
TAPE [47]	90.7	0.0	0.0	0.0	0.0	0.0
BiPE [27]	21.0	0.0	0.0	0.0	0.0	0.0
FoPE [17]	99.7	70.6	45.0	46.5	5.0	0.0
TPE (Ours)	97.2	98.2	93.9	86.6	70.9	21.1

Full NoPE baseline. Tables 25 and 26 include full NoPE, which removes positional rotation entirely. Its 32K PPL is 14945.8 and its Passkey accuracy is 0.0% from 2K onward, showing that removing all rotation alone does not recover TPE's long-range behavior. This comparison does not constitute a matched ablation of selectively unrotated low-frequency dimensions.

D.10 Checkpoint-level Variability

Tables 27 and 28 report mean and standard deviation across three checkpoints for RoPE, FoPE, and TPE at the 60M and 300M scales. Among these three methods, TPE has the lowest mean PPL at every extrapolated length from 1K to 32K. These statistics characterize checkpoint-level variation, rather than establishing independent-seed uncertainty for all model scales and evaluation metrics. The separate five-run TPE-0.3B experiment is retained in Figure 6.

Table 27: Perplexity of 60M models across three checkpoints, reported as mean ± standard deviation. Lowest mean PPL in each column is bold.

Method	**512**	**1K**	**2K**	**4K**	**8K**	**16K**	**32K**
RoPE	35.16 ± 0.05	99.31 ± 0.15	288.24 ± 2.38	518.91 ± 13.20	804.92 ± 30.51	1121.59 ± 58.89	1428.97 ± 131.21
FoPE	**35.07 ± 0.03**	34.58 ± 0.01	39.95 ± 0.15	61.97 ± 0.76	128.09 ± 2.20	259.42 ± 2.86	457.37 ± 7.31
TPE (Ours)	35.61 ± 0.06	**34.18 ± 0.06**	**35.17 ± 0.10**	**40.35 ± 0.46**	**52.22 ± 1.71**	**76.70 ± 3.77**	**127.35 ± 6.11**

Table 28: Perplexity of 300M models across three checkpoints, reported as mean ± standard deviation. Lowest mean PPL in each column is bold.

Method	**512**	**1K**	**2K**	**4K**	**8K**	**16K**	**32K**
RoPE	24.11 ± 0.00	54.06 ± 0.32	180.17 ± 3.57	462.85 ± 0.83	1072.23 ± 68.18	2395.68 ± 523.20	3914.68 ± 1018.45
FoPE	**23.99 ± 0.01**	23.38 ± 0.03	25.43 ± 0.10	34.78 ± 0.61	74.48 ± 2.02	199.21 ± 5.25	527.52 ± 41.64
TPE (Ours)	24.23 ± 0.02	**23.11 ± 0.02**	**23.18 ± 0.03**	**25.21 ± 0.22**	**31.30 ± 1.04**	**47.39 ± 3.33**	**103.20 ± 8.31**

D.11 Performance stability

Figure 6 evaluates the stability of OPT under five independent pretraining run trials and reports length extrapolation perplexity. Introducing randomness in virtual-position construction leads to measurable variation, with the best–worst gap becoming more apparent beyond 8k. Nevertheless, the fluctuation remains bounded and TPE consistently preserves a clear margin over the reference baseline: even the worst trial substantially outperforms FoPE at extreme lengths (e.g., 16k: 60.4 vs. 202.0; 32k: 135.9 vs. 540.6 perplexity), demonstrating robust long-context learning across runs.

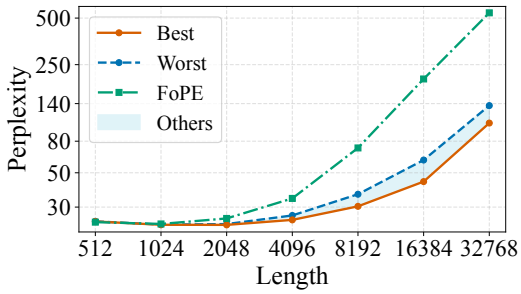
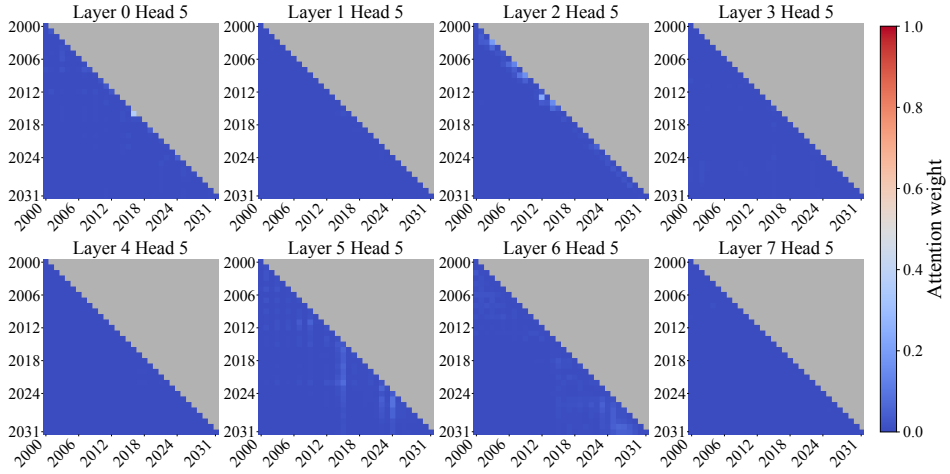


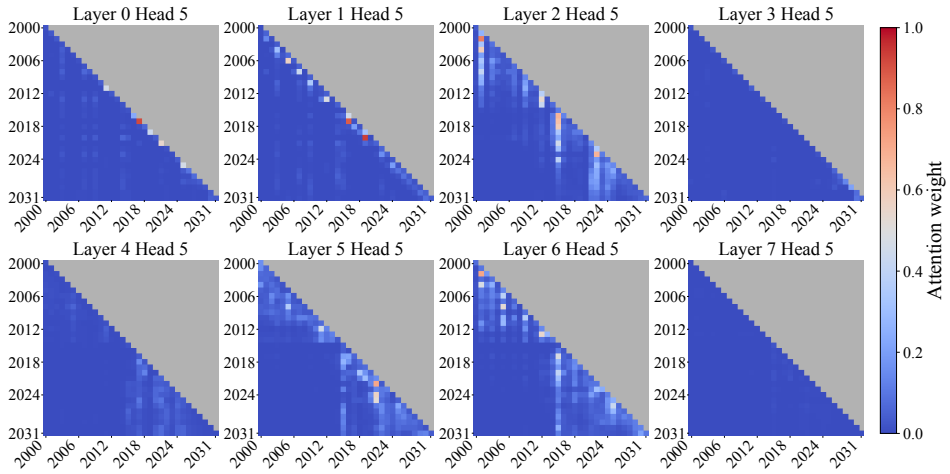
Figure 6: Perplexity under length extrapolation for TPE-0.3B across five independent pretraining run trials. The shaded region (Others) denotes the range between the best and worst runs, while the Best and Worst curves highlight the highest- and lowest-performing trials, respectively. We include FoPE as a reference baseline for comparison.

D.12 Visualization of attention weight

Building on Figure 7, the attention heatmaps reveal a qualitative difference in how RoPE and TPE model local context in the far-position regime. For RoPE, many entries close to the main diagonal are nearly zero across layers, meaning that neighboring tokens around positions 2000–2031 receive little to no attention from each query. This pattern suggests that RoPE fails to reliably encode local relationships and instead spreads attention mass more uniformly over a broad set of distant tokens, which is consistent with the rapid saturation observed in the cumulative-weight plot. In contrast, TPE exhibits a more structured distribution: attention concentrates along the diagonal with clearer local bands and occasional sparse peaks, indicating stronger locality and better positional discrimination even late in the sequence. Together with the much larger variance reported for TPE in the final layer, these visualizations support the conclusion that TPE mitigates RoPE's long-range attention homogenization and preserves useful near-neighbor signals for long-context modeling.



(a) RoPE attention weight matrices.



(b) Our TPE attention weight matrices.

Figure 7: Heatmap visualizations of attention weight matrices under different positional encodings. The plots are generated using input sequences of 2k tokens, recording the attention weights from specific layers and heads. The color intensity represents the magnitude of the corresponding attention weights.

E Limitations

Evaluation scale and data. Our experiments cover models up to 3B under a limited pretraining budget. The additional mixed-corpus experiment uses 20.05B tokens but is conducted at the 60M scale, so it does not establish larger-scale mixed-data performance. Validation on larger models, broader corpora, and longer training schedules remains future work.

Post-hoc adaptation budget. We evaluate continued-training variants with context extension from 512 to 1K. RoPE→TPE improves long-context PPL, while native TPE has stronger extreme-length retrieval under the evaluated adaptation conditions. These results demonstrate benefits of native TPE and a feasible migration path, rather than the necessity of from-scratch training or the best achievable post-hoc performance under larger budgets.

Metric range and positional trade-off. The 32K PPL improvement does not establish reliable retrieval at 32K, as we do not report Passkey retrieval results at this length for models pretrained with a 512-token context; Passkey accuracy there is close to zero. TPE also relaxes exact global shift invariance at reset boundaries. The no-reset comparison supports bounded local re-indexing within TPE, but does not establish modulo reset as uniquely optimal among alternative bounded mappings. Those alternatives and more extensive boundary-specific evaluations remain open directions.

F Broader Impacts

The development of long-context modeling methods such as TPE has broad implications for building more useful and accessible language technologies. By enabling stronger length extrapolation from short-context pretraining, TPE can help language models process longer documents, technical materials, and multi-turn interactions with improved stability. This capability may support applications in education, scientific research, software engineering, legal and medical document analysis, and knowledge-intensive assistance, where users often need models to understand and reason over long-form information.

This research aligns with the broader goal of developing AI systems that better serve human needs. By improving the efficiency and scalability of long-context modeling, TPE may reduce the computational barrier for training and deploying capable language models, making advanced long-context technologies more accessible to research communities and practical applications. We advocate for the development and deployment of such techniques in human-beneficial directions, including assisting knowledge discovery, improving productivity, supporting education, and enabling more reliable interaction with complex information.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the proposed Tree Rotary Positional Encoding (TPE), the Offset-based Positional Training (OPT) strategy, the theoretical claims, and the main empirical claims.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations in Appendix E.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The paper clearly states all theoretical results (Theorems 1, 2, and 3) along with the assumptions in Section 4.3 (Theoretical Analysis), and provides complete, step-by-step proofs for each theorem in Appendix A. All theorems, lemmas, and formulas are numbered and properly cross-referenced. The proofs explicitly list the supporting lemmas (e.g., Lemma 1)-5 and their assumptions, with rigorous logic and complete derivations. Furthermore, for the key assumption that $\mathbb{E}[a_r^2] > \mathbb{E}[b_r^2]$, we provide empirical validation at the end of Appendix A.4, which further strengthens the reliability of the theoretical assumptions.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We fully disclose all information needed to reproduce the main experimental results of the paper, see Section 5 and Appendix B.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification:The code, model checkpoints, and evaluation scripts are publicly available at `https://github.com/AIGCodeOS/aigcode_tpe`, as linked in the main paper.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide full experimental detail content in our experimental settings (see Section 5 and Appendix B).

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We evaluate statistical significance by reporting mean and standard deviation across multiple runs with different random seeds.

   Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide detailed information about on computing resources in Appendix B.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The paper meets the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

Justification: We discuss broader impacts in Appendix F.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: This paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We strictly follow the license of the assets.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.

- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We publicly release the implementation of TPE and the accompanying experimental code at `https://github.com/AIGCodeOS/aigcode_tpe`. The repository provides documentation for using the released code and reproducing the experiments.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: This paper does not involve LLMs.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.