BEYOND REAL: IMAGINARY EXTENSION OF ROTARY POSITION EMBEDDINGS FOR LONG-CONTEXT LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Rotary Position Embeddings (RoPE) have become a standard for encoding sequence order in Large Language Models (LLMs) by applying rotations to query and key vectors in the complex plane. Standard implementations, however, utilize only the real component of the complex-valued dot product for attention score calculation. This simplification discards the imaginary component, which contains valuable phase information, leading to a potential loss of relational details crucial for modeling long-range dependencies. In this paper, we propose an extension that re-incorporates this discarded imaginary component. Our method leverages the full complex-valued representation to create a dual-component attention score. We theoretically and empirically demonstrate that this approach enhances the modeling of long-context dependencies by preserving more positional information. Furthermore, evaluations on a suite of long-context language modeling benchmarks show that our method consistently improves performance over the standard RoPE, with the benefits becoming more significant as context length increases.

1 INTRODUCTION

Large Language Model (LLM) based on attention mechanism (Vaswani et al., 2017) now dominates Natural Language Processing (NLP) (OpenAI, 2023; Sun et al., 2024; OpenAI, 2024; Yang et al., 2025a), particularly in the long-context arena (Hassabis & Kavukcuoglu, 2024; Young et al., 2024; Cai et al., 2024), where attention overcomes the long-dependency bottlenecks of earlier architectures (LeCun et al., 1995; Schmidhuber et al., 1997). Recent work extends their context length to the million-token scale (Liu et al., 2024a; InternLM, 2025), and the key driver is position-embedding design (Su et al., 2024; Press et al., 2022; Peng et al., 2024). Among current LLMs, Rotary Position Embedding (RoPE) (Su et al., 2024) has become the canonical choice (Dubey et al., 2024; Meta, 2024a;b). It encodes the absolute position of every query and key vector $\boldsymbol{q}_t, \boldsymbol{k}_s$, namely token indices s, t with a rotary matrix or complex multiplication, and when the two vectors make a dot product, it injects their relative position $t - s$, namely the relative distance, into the attention scores, thus combining the merits of traditional absolute and relative position embeddings (Vaswani et al., 2017; Dai et al., 2019; Yan et al., 2019) and securing widespread adoption.

Nevertheless, RoPE also has notable shortcomings, including poor length extrapolation (Press et al., 2022; Chen et al., 2023; bloc97, 2023), lack of data-sensitivity (Golovneva et al., 2024; Yang et al., 2025b), and no design for heterogeneous multi-modal input (Su, 2024a), prompting extensive research into its improvement. Most efforts concentrate on refining RoPE through interpolation designs (Peng et al., 2024; Liu et al., 2024c; Su, 2023), data-awareness (Zheng et al., 2024a;b), and feature-dimension partitioning (Wang et al., 2024; Wei et al., 2025). However, few work revisits the intrinsic computation of RoPE or analyze its inherent limitations (Hua et al., 2024; Dai et al., 2025). Re-examining RoPE in its complex-multiplication form reveals that the standard implementation keeps only the real part of the resulting complex attention score and discards the imaginary part outright (Su et al., 2024). Although taking the real part preserves the direct equivalence between complex multiplication and vector rotation, it incurs an irreversible information loss.

A closer look at the imaginary attention, strictly, the negative imaginary part of attention, shows that, compared with the real attention exhibiting stronger semantic locality, the imaginary heads attend more to long-context information as shown in Figure 1, promising gains on long-context tasks. Moreover, adding imaginary attention also exposes $\boldsymbol{q}_t, \boldsymbol{k}_s$ to a wider positional information range,

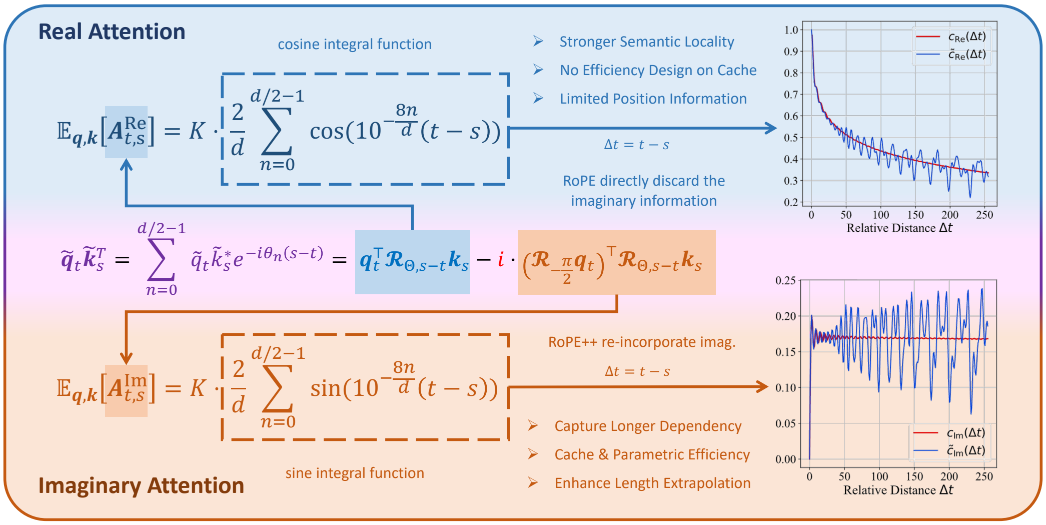

Real Attention
cosine integral function
$\mathbb{E}_{q,k}[A^{\mathrm{Re}}_{t,s}] = K \cdot \frac{2}{d}\sum_{n=0}^{d/2-1}\cos(10^{-\frac{8n}{d}}(t-s))$
Stronger Semantic Locality
No Efficiency Design on Cache
Limited Position Information
$\Delta t = t - s$
RoPE directly discard the imaginary information
$\tilde{q}_t\tilde{k}_s^T = \sum_{n=0}^{d/2-1}\tilde{q}_t\tilde{k}_s^* e^{-i\theta_n(s-t)} = q_t^\top \mathcal{R}_{\Theta,s-t}k_s - i \cdot (\mathcal{R}_{-\frac{\pi}{2}}q_t)^\top \mathcal{R}_{\Theta,s-t}k_s$
RoPE++ re-incorporate imag.
$\Delta t = t - s$
$\mathbb{E}_{q,k}[A^{\mathrm{Im}}_{t,s}] = K \cdot \frac{2}{d}\sum_{n=0}^{d/2-1}\sin(10^{-\frac{8n}{d}}(t-s))$
Capture Longer Dependency
Cache & Parametric Efficiency
Enhance Length Extrapolation
sine integral function
Imaginary Attention
$c_{\mathrm{Re}}(\Delta t)$
$\tilde{c}_{\mathrm{Re}}(\Delta t)$
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0
50
100
150
200
250
Relative Distance Δt
$c_{\mathrm{Im}}(\Delta t)$
$\tilde{c}_{\mathrm{Im}}(\Delta t)$
0.25
0.20
0.15
0.10
0.05
0.00
0
50
100
150
200
250
Relative Distance Δt

Figure 1: Overview of RoPE++. RoPE retains only the real part of the complex-valued attention score, whereas RoPE++ exploits the full complex representation to produce both real and imaginary attention. The real attention exhibits stronger semantic locality, while the imaginary attention preferentially captures long-context dependencies. RoPE++ combines the two, yielding multiple advantages.

implicitly improving length extrapolation. Therefore, we propose **RoPE++**, as illustrated in Figure 1, which re-injects the discarded imaginary component as a new group of attention heads computed in parallel with the real attentions. Particularly, we introduce **RoPE++$_{\mathbf{EH}}$** that keeps equal attention head number while halving QKV parameters as well as KV cache, and **RoPE++$_{\mathbf{EC}}$** that keeps equal cache size and doubles the number of attention heads. Theoretical analysis and pre-training experiments validate the above advantages. Both RoPE++$_{\mathrm{EH}}$ and RoPE++$_{\mathrm{EC}}$ outperform vanilla RoPE and other position embeddings on general tasks. On long-context benchmarks, RoPE++$_{\mathrm{EH}}$ achieves comparable results with vanilla RoPE with half the cache, whereas RoPE++$_{\mathrm{EC}}$ outperforms significantly at the same cache cost. Our contributions can be summarized as follows:

- We first identify the loss of imaginary information in standard RoPE and find it advantageous for capturing long-context dependencies and enhancing length extrapolation by analyzing the properties of imaginary attention.
- Building on this, we propose RoPE++, which reintroduces the imaginary computation into attention in two configurations, RoPE++$_{\mathrm{EH}}$ with equal head number and halved KV cache, and RoPE++$_{\mathrm{EC}}$ with equal cache size and doubled attention heads. Both preserve the unified absolute–relative position-embedding format.
- Pre-training and evaluation at 376M and 776M sizes show that RoPE++$_{\mathrm{EH}}$ and RoPE++$_{\mathrm{EC}}$ outperform vanilla RoPE and other position embeddings on average across short- and long-context benchmarks. Further analysis reveals that the imaginary attentions play a dominant role in modeling long-context dependencies, confirming the effectiveness of introducing imaginary attention for improved long-context capability.

2 Related Work

Rotary Position Embedding (RoPE) is the dominant position embedding in current LLMs (Dubey et al., 2024; Meta, 2024a;b; Yang et al., 2025a). We analyze its good properties in Appendix B, including unifying relative and absolute information via rotation matrices and complex multiplication, and semantic aggregation as well as long-context decay. Yet it still faces many other challenges, attracting a great deal of effort to its improvement as mentioned above. A large body of work targets length extrapolation, scaling the rotary base (bloc97, 2023; Liu et al., 2024c; Xiong et al., 2024), interpolating or compressing index ranges (Press et al., 2022; Peng et al., 2024; Jin et al., 2024), or coupling RoPE with sparse attention (Lu et al., 2024; Xiao et al., 2024; Liu et al., 2024b) to let models process contexts far longer than the training window. Other efforts extend RoPE to heterogeneous, cross-modal inputs (Su, 2024a), especially text–video sequences (Wang et al., 2024; Wei et al., 2025). Parallel lines design parametric schemes that encode contextual cues (Golovneva et al., 2024; Zheng et al., 2024a; Lin et al., 2025), refining or replacing RoPE to yield data-dependency.

However, few works revisit RoPE's intrinsic computation or analyze its inherent limitations (Hua et al., 2024; Yang et al., 2025b; Dai et al., 2025). Particularly, the imaginary information loss of RoPE in rotation format compared with the complex multiplication format remains overlooked. Although prior work has tried to incorporate the full complex computation into the self-attention mechanism or neural networks (Wang et al., 2025; Lee et al., 2022), the characteristics and functionality of the imaginary component in position embedding remain unexplored. Therefore, we propose RoPE++ and close this gap through a deep analysis of the mathematical properties of imaginary attention and extensive validation on both short- and long-context downstream tasks.

3 Methodology

We begin our method by revisiting the complex form of RoPE. Only the real part of the complex product is retained, and the imaginary part is discarded, as shown in Equation 1. Although current LLMs perform well with this real-only attention, omitting the imaginary component may remove physical information. LLM no longer sees the full magnitude and phase of the complex attention result. This raises the question: can the imaginary part be re-incorporated into the attention computation?

$$\begin{aligned}\boldsymbol{A}_{t,s} &= \mathrm{Re}\left[\sum_{n=0}^{d/2-1} \tilde{q}_t^{(n)} \tilde{k}_s^{(n)*} e^{-i\theta_n(t-s)}\right] = \mathrm{Re}\left[\sum_{n=0}^{d/2-1} \left(\tilde{q}_t^{(n)} e^{-i\theta_n t}\right)\left(\tilde{k}_s^{(n)} e^{-i\theta_n s}\right)^*\right] \\ &= \sum_{n=0}^{d/2-1} \begin{array}{l}\left(q_t^{(2n)} k_s^{(2n)} + q_t^{(2n+1)} k_s^{(2n+1)}\right) \cos\theta_n(t-s) + \\ \left(q_t^{(2n)} k_s^{(2n+1)} - q_t^{(2n+1)} k_s^{(2n)}\right) \sin\theta_n(t-s)\end{array}\end{aligned} \tag{1}$$

In this section, we will first propose our RoPE++ by re-introducing the imaginary information, in Section 3.1, as a new group of attention heads, namely imaginary attentions, compared with original real attentions. We then analyze the strengths from three aspects, the imaginary heads' stronger capture of long-context dependencies in Section 3.2, the cache and parameter reduction by combining imaginary and real heads in Section 3.3, and the impact on length extrapolation in Section 3.4.

3.1 Imaginary Extension of RoPE

We first recover the imaginary part that is discarded in Equation 1. The resulting expression is given in Equation 2. Strictly speaking, it is the negative imaginary part, and the reason will be detailed in Section 3.2. Similar to the real part, the imaginary part carries relative position information between $\boldsymbol{q}_t, \boldsymbol{k}_s$, so the formula can be rearranged into a vector form as shown in Equation 2.

$$\begin{aligned}\boldsymbol{A}_{t,s}^{\mathrm{Im}} &= -\mathrm{Im}\left[\sum_{n=0}^{d/2-1} \tilde{q}_t^{(n)} \tilde{k}_s^{(n)*} e^{-i\theta_n(t-s)}\right] = -\mathrm{Im}\left[\sum_{n=0}^{d/2-1} \left(\tilde{q}_t^{(n)} e^{-i\theta_n t}\right)\left(\tilde{k}_s^{(n)} e^{-i\theta_n s}\right)^*\right] \\ &= \sum_{n=0}^{d/2-1} \begin{array}{l}\left(q_t^{(2n)} k_s^{(2n)} + q_t^{(2n+1)} k_s^{(2n+1)}\right) \sin\theta_n(t-s) - \\ \left(q_t^{(2n)} k_s^{(2n+1)} - q_t^{(2n+1)} k_s^{(2n)}\right) \cos\theta_n(t-s)\end{array}\end{aligned} \tag{2}$$

We observe that the imaginary attention still follows a rotation form and can be decomposed into absolute position embeddings on $\boldsymbol{q}_t, \boldsymbol{k}_s$, as shown in Equation 3. Specifically, the embedding applied to $\boldsymbol{k}_s$ is identical to that used in the real attention in Equation 6 in Appendix B. For $\boldsymbol{q}_t$, the embedding is equivalent to rotating the vector by $-\pi/2$ before applying the same embedding in the real case.

$$\begin{aligned}\boldsymbol{A}^{\text{Im}}_{t,s} &= \underbrace{\sum_{n=0}^{d/2-1} \begin{bmatrix} q_t^{(2n+1)} \\ -q_t^{(2n)} \end{bmatrix}^\top \begin{bmatrix} \cos\theta_n(t-s) & \sin\theta_n(t-s) \\ -\sin\theta_n(t-s) & \cos\theta_n(t-s) \end{bmatrix} \begin{bmatrix} k_s^{(2n)} \\ k_s^{(2n+1)} \end{bmatrix}}_{\text{Relative PE}} \\ &= \underbrace{\sum_{n=0}^{d/2-1} \left(\begin{bmatrix} \cos\theta_n t & -\sin\theta_n t \\ \sin\theta_n t & \cos\theta_n t \end{bmatrix} \begin{bmatrix} q_t^{(2n+1)} \\ -q_t^{(2n)} \end{bmatrix} \right)^\top \left(\begin{bmatrix} \cos\theta_n s & -\sin\theta_n s \\ \sin\theta_n s & \cos\theta_n s \end{bmatrix} \begin{bmatrix} k_s^{(2n)} \\ k_s^{(2n+1)} \end{bmatrix} \right)}_{\text{Absolute PE}} \end{aligned} \tag{3}$$

We thus obtain an expression for the imaginary attention, strictly speaking, the negative imaginary attention. If we denote the rotation matrix as $\boldsymbol{\mathcal{R}}_{\cdot}$ and $\boldsymbol{\mathcal{R}}_{\Theta,\cdot}$. The latter is parameterized with $\theta_0, \cdots, \theta_{d/2-1}$. The computation of real and imaginary attention can be summarized in Equation 4.

$$\begin{aligned} \boldsymbol{A}^{\text{Re}}_{t,s} &= \text{Re}\left[\sum_{n=0}^{d/2-1} \tilde{q}_t^{(n)} \tilde{k}_s^{(n)*} e^{i\theta_n(s-t)}\right] = \left(\boldsymbol{\mathcal{R}}_{\Theta,t}\boldsymbol{q}_t\right)^\top \boldsymbol{\mathcal{R}}_{\Theta,s}\boldsymbol{k}_s = \boldsymbol{q}_t^\top \boldsymbol{\mathcal{R}}_{\Theta,s-t}\boldsymbol{k}_s \\ \boldsymbol{A}^{\text{Im}}_{t,s} &= -\text{Im}\left[\sum_{n=0}^{d/2-1} \tilde{q}_t^{(n)} \tilde{k}_s^{(n)*} e^{i\theta_n(s-t)}\right] = \left(\boldsymbol{\mathcal{R}}_{\Theta,t}\boldsymbol{\mathcal{R}}_{-\frac{\pi}{2}}\boldsymbol{q}_t\right)^\top \boldsymbol{\mathcal{R}}_{\Theta,s}\boldsymbol{k}_s = \left(\boldsymbol{\mathcal{R}}_{-\frac{\pi}{2}}\boldsymbol{q}_t\right)^\top \boldsymbol{\mathcal{R}}_{\Theta,s-t}\boldsymbol{k}_s \end{aligned} \tag{4}$$

Notably, the newly introduced imaginary component retains the key property of the original RoPE, that it can still be formulated either as a relative position or as an absolute position embedding. The only required adjustment is to rotate $\boldsymbol{q}_t$ by $-\pi/2$ and then apply the standard position embedding to obtain the imaginary term. We refer to RoPE augmented with this imaginary extension as **RoPE++**. This augmentation raises further questions: what semantics does the imaginary attention convey, does it introduce additional overhead, and can it enhance model performance?

3.2 Capture Longer Dependency

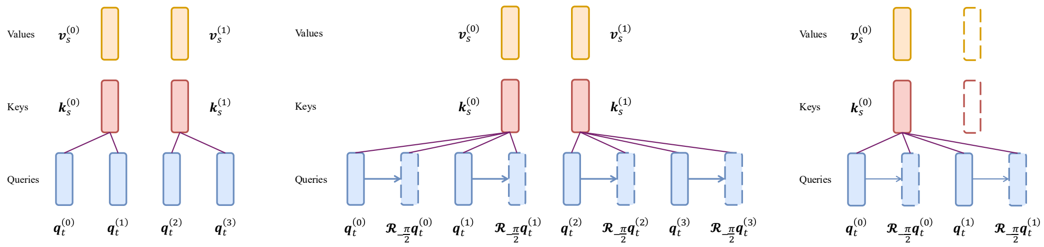
As stated in Preliminary in Appendix B, the original RoPE-based attention or real attention exhibits semantic aggregation and long-context decay, both governed by its characteristic curve, as shown in Equation 7 and Figure 1. Similarly, we can derive the characteristic curve for the imaginary attention in RoPE++. It is the average of $\sin(\theta\Delta t)$ over the same frequency distribution, approximating a sine integral function as shown in Equation 5 and Figure 1.

$$c_{\text{Im}}(\Delta t) = \frac{2}{d}\sum_{n=0}^{d/2-1} \sin\left(10^{-\frac{8n}{d}}\Delta t\right), \quad \tilde{c}_{\text{Im}} = \int_{10^{-4}}^{1} \frac{\sin\theta t}{\theta \ln 10^4} \mathrm{d}\theta = \text{Si}(\Delta t) - \text{Si}\left(\frac{\Delta t}{10^4}\right) \tag{5}$$

Although modeling distance with $\sin(\theta\Delta t)$ is counter-intuitive, since $\sin(\theta\Delta t)$ is zero at zero relative distance, rises, then falls, unlike $\cos(\theta\Delta t)$'s monotonic drop in the first half-period, the characteristic curve of the imaginary attention still shares the semantic-aggregation property of the real part. For $\Delta t > 0$, when $\boldsymbol{q}_t, \boldsymbol{k}_s$ are similar, their attention is on average larger regardless of relative distance, which is the reason why we take the negative imaginary part as imaginary attention. Moreover, on average, this component attends more to distant positions. As shown in Figure 1, its characteristic curve declines very slowly beyond a certain distance. Consequently, the imaginary part assigns more weight to the long-context region than the real part, helping LLM retrieve long-context information.

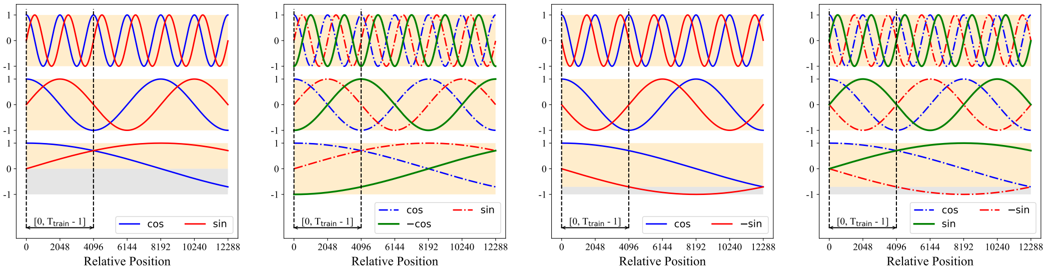
3.3 Cache and Parametric Efficiency

As described earlier, computing the imaginary attention requires only rotating the $\boldsymbol{q}_t$ by $-\pi/2$, while every other operation is identical to the original RoPE. Because the positional embedding of $\boldsymbol{k}_s$ is unchanged, we can interleave the $-\pi/2$-rotated $\boldsymbol{q}_t$ with the original $\boldsymbol{q}_t$ and perform the real and imaginary attention in a single pass in FlashAttention (Dao, 2024). Consequently, no extra KV cache is introduced, and the method plugs directly into MHA or GQA (Ainslie et al., 2023), merely


Values
Keys
Queries
$v_s^{(0)}$
$v_s^{(1)}$
$k_s^{(0)}$
$k_s^{(1)}$
$q_t^{(0)}$
$q_t^{(1)}$
$q_t^{(2)}$
$q_t^{(3)}$
$\mathcal{R}_{-\frac{\pi}{2}}q_t^{(0)}$
$\mathcal{R}_{-\frac{\pi}{2}}q_t^{(1)}$
$\mathcal{R}_{-\frac{\pi}{2}}q_t^{(2)}$
$\mathcal{R}_{-\frac{\pi}{2}}q_t^{(3)}$

(a) GQA with RoPE (b) GQA with RoPE++$_{\text{EC}}$ (c) GQA with RoPE++$_{\text{EH}}$

Figure 2: Visualization of GQA with different RoPE schema. RoPE++$_{\text{EC}}$ shares equal cache and twice the attention head with RoPE, while RoPE++$_{\text{EH}}$ has equal attention head and half the KV cache.



(a) Position embedding of $q^{(2n)}, k^{(2n+1)}$ in RoPE (b) Position embedding of $q^{(2n)}, k^{(2n+1)}$ in RoPE++ (c) Position embedding of $q^{(2n+1)}, k^{(2n)}$ in RoPE (d) Position embedding of $q^{(2n+1)}, k^{(2n)}$ in RoPE++

Figure 3: Comparison of trained position embedding interval between RoPE and RoPE++. The area within the dashed line represents trained relative position, and that beyond is in length extrapolation, with learned position embedding values colored in yellow and the opposite in gray.

doubling the attention head group size, as shown in Figure 2b. We refer to this configuration as **RoPE++$_{\text{EC}}$**, namely RoPE++ with equal cache size. The only cost of RoPE++$_{\text{EC}}$ is an additional imaginary attention computed alongside the real one under the fixed QKV parameter budget.

Conversely, if the total head number is kept fixed, both QKV parameters and KV cache sizes are halved. We refer to this configuration as **RoPE++$_{\text{EH}}$**, namely RoPE++ with equal attention head number, as shown in Figure 2c. In long-context scenarios, RoPE++$_{\text{EH}}$ halves the cache and raises throughput. Experiments in Section 4 demonstrate that RoPE++$_{\text{EC}}$ outperforms the original RoPE, especially on long-context tasks, and RoPE++$_{\text{EH}}$ delivers comparable or even superior results.

Importantly, the imaginary and real attention, though computed independently and treated as separate heads, must share the same parameter. Allocating distinct subsets of heads to imaginary and real attention would keep head number and cache size unchanged and appear to permit a fair comparison with the original RoPE-based LLM, but it would effectively collapse back to standard RoPE, since rotating $\boldsymbol{q}_t$ in imaginary attention by $\pi/2$ yields real attention, with no architecture modification.

3.4 Impact on Length Extrapolation

A closer inspection of the real and imaginary attention computations reveals an interesting discovery. In vanilla RoPE-based attention, or real attention, as shown in Equation 6, even-index query dimensions $\boldsymbol{q}^{(2n)}$ and odd-index key dimensions are multiplied only by $\cos\theta_n(t-s)$ and $\sin\theta_n(t-s)$ whose values are always non-negative when θ_n is small. Once the input length exceeds the pre-training context length, these dimensions encounter out-of-distribution (OOD) negative embeddings as shown in Figure 5f and thus extrapolate poorly (Liu et al., 2024c; Peng et al., 2024). In RoPE++ as shown in Equation 3, these dimensions are multiplied by $-\cos\theta_n(t-s)$ and $\sin\theta_n(t-s)$ in the imaginary attention, so during pre-training, they have already observed both negative and positive position embedding as well as their maximum and minimum value ± 1. Consequently, these dimensions no longer suffer from the length extrapolation problem in longer contexts (Liu et al., 2025b).

Likewise, odd-index query dimensions $\boldsymbol{q}^{(2n+1)}$ and even-index key dimensions $\boldsymbol{k}^{(2n)}$ encounter only $\cos\theta_n(t-s)$ and $-\sin\theta_n(t-s)$ in the real attention, and the imaginary attention further exposes them to $\cos\theta_n(t-s)$ and $\sin\theta_n(t-s)$. Yet this alone does not expand the position embedding range trained in pre-training, as shown in Figure 5h and Figure 5j. However, when real and imaginary attention are combined, $\boldsymbol{q}_t, \boldsymbol{k}_s$ in RoPE++ attains the full cos and sin value range, once the training length exceeds half the sinusoidal period, whereas the vanilla RoPE requires a full period. Consequently, more dimensions in RoPE++ observe complete positional information. Therefore, perplexity grows more slowly beyond the maximum supported context length (Liu et al., 2024c; Men et al., 2024).

4 EXPERIMENT

4.1 SETUP

We validate RoPE++ at both 776M and 376M model sizes, with architectural details in Appendix C. Both models are pre-trained on DCLM-Baseline-1.0 corpus (Li et al., 2024) by HuggingFace Transformers (Wolf et al., 2020) on 8 NVIDIA H200 160 GB GPUs. For each size, we use a batch size of 0.5M tokens and pre-train for 50B tokens. We use AdamW (Loshchilov et al., 2017) optimizer with weight decay 0.1, a maximum learning rate of 5e-4, and a warmup-stable-decay scheduler. We use the first 0.5B tokens for warmup, and the final 5B tokens for decay, and the learning rate ends at 0.

We compare our RoPE++ with standard RoPE (Su et al., 2024) and other well-known position embedding designs, including FoPE (Hua et al., 2024), NoPE (Haviv et al., 2022), as well as ALiBi (Press et al., 2022). We pre-train RoPE, RoPE++, and FoPE on 4k context length with an initial rotary base of 10000. Since ALiBi and NoPE train unstably at 4k, as we have tried, we train them on 1k context length while keeping the batch size the same. For RoPE and RoPE++, we conduct continuous pre-training on long-context. Following Xiong et al. (2024); Lv et al. (2024), we scale the rotary base from 10000 to 500000 and train for 10B additional tokens from DCLM on 32k context length, using a cosine-annealing learning-rate scheduler and leaving all other settings unchanged.

4.2 SHORT-CONTEXT EVALUATION

We evaluate both short-context and long-context tasks based on OpenCompass (Contributors, 2023). For short-context evaluation, we measure perplexity on WikiText (Merity et al., 2017) and LAMBADA(Paperno et al., 2016) and assess downstream tasks mainly in Open LLM Leaderboard (HuggingFace, 2023), including TruthfulQA (Lin et al., 2022), PIQA (Bisk et al., 2020), HellaSwag (Zellers et al., 2019), WinoGrande (Sakaguchi et al., 2020), ARC-e (Clark et al., 2018), GPQA (Rein et al., 2023), SocialIQA (Sap et al., 2019), OpenBookQA (Mihaylov et al., 2018), and SuperGLUE (Wang et al., 2019). Specifically, ALiBi and NoPE are tested within a 1k context length, while others are tested with 4k. Since the data length is generally less than 1k, this comparison is still fair.

The results are shown in Table 1. Our RoPE++$_{\text{EC}}$ and RoPE++$_{\text{EH}}$ achieve the best average scores on short-context tasks compared with RoPE and every other position embedding design. Notably, RoPE++$_{\text{EH}}$ surpasses standard RoPE with only half the KV-cache and QKV parameters. After continuous long-context pre-training, RoPE++ still retains this edge over RoPE on short-text benchmarks.

4.3 LONG-CONTEXT EVALUATION

For long-context evaluation, we evaluate downstream performance at varying lengths with the classical synthetic benchmarks, RULER (Hsieh et al., 2024) and BABILong (Kuratov et al., 2024). The results are shown in Table 2 and Figure 6. We highlight the comparison with RoPE in long-context training because RoPE is the position embedding currently used by nearly all long-context LLMs, and its long-context training pipeline is already highly mature.

On RULER and BABILong up to 64k context, our RoPE++ again acquires the highest scores. Particularly, RoPE++$_{\text{EH}}$ achieves comparable performance with vanilla RoPE using half the KV-cache and QKV parameters, while RoPE++$_{\text{EC}}$ delivers significant gains at the same cache size. Although RoPE occasionally edges ahead at a few shorter context lengths, RoPE++, including both

	Wiki ppl ↓	**LMB** ppl ↓	**TQA** acc ↑	**PIQA** acc ↑	**Hella** acc ↑	**Wino** acc ↑	**ARC-e** acc ↑	**GPQA** acc ↑	**SIQA** acc ↑	**OBQA** acc ↑	**SG** acc ↑	**Avg.**
376M Short												
RoPE	19.9	32.7	35.5	66.3	34.8	50.9	39.3	24.8	38.6	**27.4**	43.7	40.1
FoPE	**19.3**	33.0	33.8	65.9	34.5	**53.0**	37.0	**28.8**	39.5	24.2	43.6	40.0
NoPE	31.4	43.7	35.5	62.2	28.7	51.1	34.7	25.3	37.0	25.8	**46.2**	38.5
ALiBi	31.2	44.7	35.2	62.8	29.2	50.7	36.2	26.8	35.8	22.8	45.9	38.4
RoPE++$_{\text{EH}}$	20.8	33.6	36.3	66.4	34.5	52.5	40.9	23.7	**40.5**	24.8	43.2	40.3
RoPE++$_{\text{EC}}$	19.4	**32.6**	**37.3**	**68.0**	**35.6**	**53.0**	**41.3**	25.8	40.3	23.2	44.8	**41.0**
376M Long												
RoPE	20.4	**33.8**	35.4	64.9	34.1	50.6	40.4	21.2	**39.4**	27.4	43.5	39.6
RoPE++$_{\text{EH}}$	21.7	34.8	35.2	64.5	**34.3**	49.9	**41.5**	**22.7**	40.0	27.0	43.1	39.8
RoPE++$_{\text{EC}}$	**20.0**	33.9	**37.1**	**66.1**	34.1	**53.4**	38.1	21.2	39.2	**28.4**	**43.7**	**40.1**
776M Short												
RoPE	14.8	27.3	35.5	**70.1**	**43.7**	52.3	43.4	25.8	41.3	21.8	43.6	42.0
FoPE	**14.7**	**27.1**	33.6	68.7	43.4	52.9	**45.0**	24.8	39.7	24.8	45.4	42.0
NoPE	19.5	33.4	34.5	67.1	37.1	51.1	39.9	25.3	38.3	27.4	44.7	40.6
ALiBi	21.6	34.9	33.3	66.8	35.2	49.5	41.6	26.3	37.8	27.0	**46.0**	40.4
RoPE++$_{\text{EH}}$	15.6	28.1	35.4	69.6	42.7	**53.5**	**45.0**	15.8	**41.6**	26.8	42.4	42.5
RoPE++$_{\text{EC}}$	14.8	27.3	**36.1**	69.3	43.6	52.3	43.7	**28.3**	40.1	**27.6**	44.4	**42.8**
776M Long												
RoPE	14.6	27.3	35.1	68.9	43.1	51.5	**47.6**	21.7	40.7	20.2	42.6	41.3
RoPE++$_{\text{EH}}$	15.3	28.1	**35.4**	69.9	41.9	**52.6**	43.2	28.3	**41.0**	22.2	43.4	42.0
RoPE++$_{\text{EC}}$	**14.4**	**27.1**	35.2	**70.4**	**43.7**	**52.6**	44.8	**31.8**	40.8	**27.6**	**44.3**	**43.5**

Table 1: Results on short-context tasks for 776M and 376M models pre-trained in short context and further trained on 32k context length. Best results are highlighted in bold, while the second-best results are underlined for broader comparison. Our RoPE++ achieves the best performance on average among different model sizes and training settings.

	RULER						**BABILong**							
	4k	8k	16k	32k	64k	Avg.	0k	2k	4k	8k	16k	32k	64k	Avg.
376M Long														
RoPE	31.6	25.6	22.0	9.5	5.5	18.8	23.3	17.7	16.1	9.1	9.4	5.9	7.8	12.8
RoPE++$_{\text{EH}}$	29.9	28.4	17.6	9.4	5.9	18.2	29.6	14.1	15.6	12.2	9.9	8.3	9.7	14.2
RoPE++$_{\text{EC}}$	**36.1**	**33.0**	**29.1**	**17.7**	**9.0**	**25.0**	**27.2**	**19.8**	**19.8**	**16.1**	**15.8**	**12.3**	**12.8**	**17.7**
776M Long														
RoPE	37.4	35.1	33.0	21.2	10.4	27.4	14.8	**33.5**	**30.7**	23.6	22.0	15.1	12.1	21.7
RoPE++$_{\text{EH}}$	38.7	35.4	**33.8**	**24.6**	10.7	28.6	**33.6**	31.9	26.5	18.6	16.2	11.0	12.2	21.4
RoPE++$_{\text{EC}}$	**42.7**	**38.6**	33.4	21.7	**10.9**	**29.4**	22.1	32.4	29.9	**24.4**	**24.5**	**18.6**	**14.8**	**23.8**

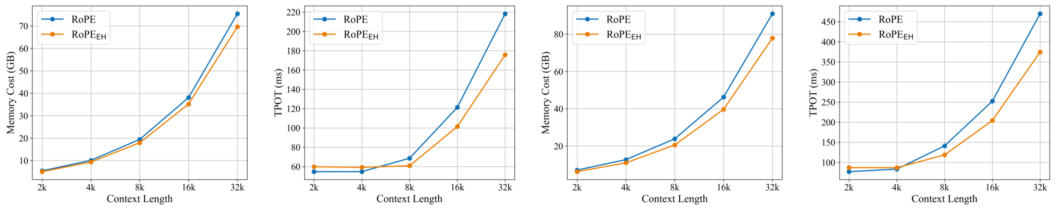
Table 2: Results on long-context tasks, including RULER and BABILong for 776M and 376M models further trained with 5B tokens in 32k context length. Best results are highlighted in bold. Our RoPE++ achieves the best performance on average, especially in long-context scenarios.

RoPE++$_{\text{EC}}$ and RoPE++$_{\text{EH}}$, maintains more stable performance as context length grows and achieves best performance in 64k context length extrapolation consistently.

5 DISCUSSION

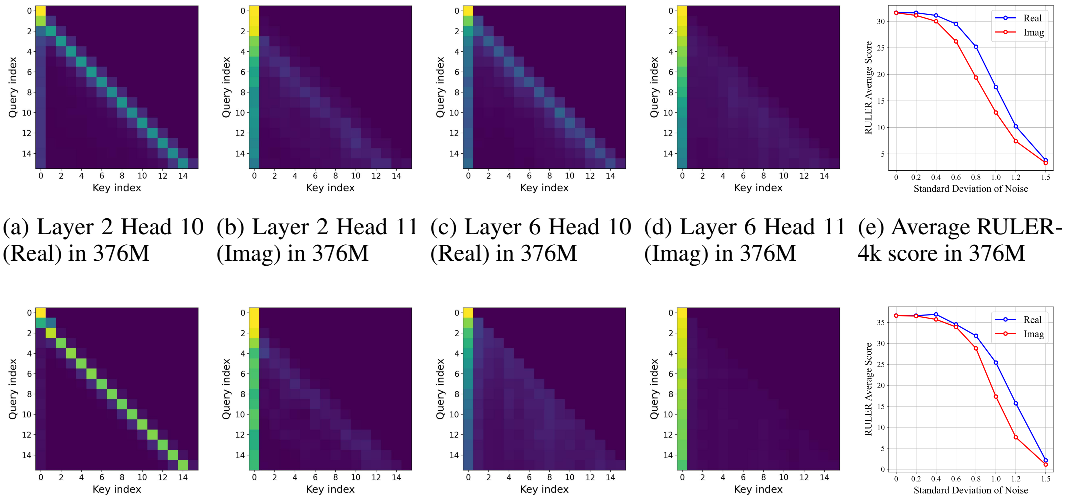
5.1 ROPE++ AS CACHE OPTIMIZATION

As mentioned in Section 3.3, RoPE++$_{\text{EH}}$ halves KV cache and QKV parameters while keeping the attention head number equal, yielding evident efficiency gains. We validate this efficiency strength by assessing the memory cost as well as Time-Per-Output-Token (TPOT) of 376M and 776M models, from 2k to 32k context length. We conduct the efficiency evaluation on a single NVIDIA H200



(a) Memory Cost of 376M (b) TPOT of 376M (c) Memory Cost of 776M (d) TPOT of 776M

Figure 4: Efficiency comparison between RoPE and RoPE++$_{\text{EH}}$ in 376M and 776M model. RoPE++$_{\text{EH}}$ lowers memory cost and accelerates decoding, and the margin widens as context grows.



(a) Layer 2 Head 10 (Real) in 376M (b) Layer 2 Head 11 (Imag) in 376M (c) Layer 6 Head 10 (Real) in 376M (d) Layer 6 Head 11 (Imag) in 376M (e) Average RULER-4k score in 376M

(f) Layer 5 Head 10 (Real) in 776M (g) Layer 5 Head 11 (Imag) in 776M (h) Layer 11 Head 10 (Real) in 776M (i) Layer 11 Head 11 (Imag) in 776M (j) Average RULER-4k curve in 776M

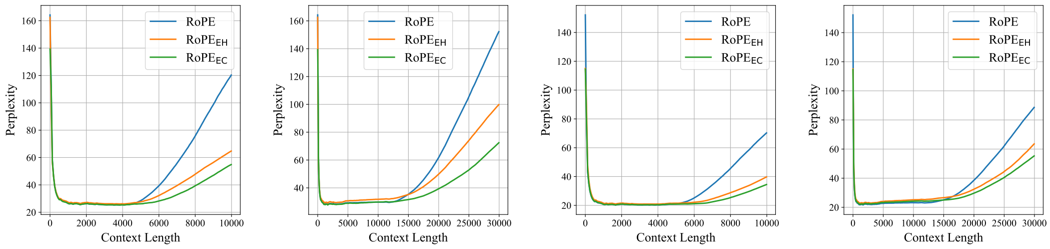
Figure 5: Attention-score patterns and long-context performance in 376M and 776M RoPE++ models. Imaginary heads attend markedly to global information, whereas real heads focus more on local context. Adding Gaussian noise to imaginary attention degrades long-context performance more severely, over 8 points, than the same perturbation applied to real attention.

160BG GPU, with a batch size of 8 samples. The results are shown in Figure 4. At both 376M and 776M, RoPE++$_{\text{EH}}$ consistently reduces memory cost and speeds up decoding, with the margin widening as context length increases.

5.2 Attention Pattern of RoPE++

To verify how imaginary attention captures long-context dependencies and to contrast it with real attention in RoPE++, we inspect the attention patterns of short-context-trained RoPE++$_{\text{EC}}$ at 376M and 776M as shown in Figure 5. Odd-index imaginary attention highlights the initial positions more strongly than even-index real heads, indicating a stronger global focus. Since prior work (Liu et al., 2025a; Wei et al., 2025) shows that dimensions attending globally are more critical for long-context semantics, imaginary attention may play the dominant role in long-context tasks.

For further verification, we design the following validation experiment. We add Gaussian noise with equal standard deviation to the imaginary and real attention components separately, and monitor the change in RoPE++ performance on long-context tasks, such as the average score of RULER-4k. Curves for RULER-4k versus standard deviation are plotted for both real and imaginary attention. When the standard deviation σ is small ($\sigma < 0.2$), scores with corrupted real or imaginary attentions stay close to the baseline; when it is large enough ($\sigma = 1.5$), both drop sharply. Importantly, in the intermediate range, adding noise to the imaginary attention always performs worse than corrupting



(a) Perplexity curve of pre-trained 376M.
(b) Perplexity curve of NTK-scaled 376M.
(c) Perplexity curve of pre-trained 776M.
(d) Perplexity curve of NTK-scaled 776M.

Figure 6: Perplexity comparison between RoPE and RoPE++ in 376M and 776M models.

the real part. When $\sigma = 1.0$, for example, the real-noised RoPE++ outperforms the imaginary-noised one by 5 points at 376M and 8 points at 776M, which demonstrates a significant gap. Thus, impairing the imaginary heads degrades long-context performance more, confirming that imaginary attention plays a more dominant role in long context modeling.

5.3 Perplexity Curve of RoPE++

Length extrapolation is a central issue for long-context LLMs. We have already shown that RoPE++ outperforms RoPE on long-context downstream tasks in training-based length extrapolation. However, RoPE++ cannot directly extrapolate like FoPE (Hua et al., 2024) or PaTH (Yang et al., 2025b). Once the inference exceeds the maximum supported context length, perplexity begins to rise. Interestingly, as discussed in Section 3.4, every even-index dimension in query vectors and odd-index dimension in key vectors are trained with full value range of position embeddings, and every dimension has seen both positive and negative positions during training. Consequently, the perplexity curve of RoPE++ climbs more gradually (Liu et al., 2024c).

This is verified in Figure 6, where we compare the perplexity of short-context-trained RoPE and RoPE++ on 376M and 776M model sizes. With or without fixed-NTK interpolation based on scaling factor $\lambda = 4$, both curves rise at the same context length with RoPE, indicating an identical stable context upper bound. Beyond that point, however, RoPE++'s perplexity increases more slowly, confirming the earlier prediction about its extrapolation behavior.

Limitation

As noted above, RoPE++ markedly boosts performance on both short- and long-context tasks, yet it does not deliver plug-and-play length extrapolation and still falls behind such extrapolation designs as FoPE and PaTH. Nevertheless, as a method that reintroduces imaginary attention, raising performance under fixed memory or improving efficiency while preserving accuracy, RoPE++ can be integrated with part of those designs. Additionally, thanks to the oddity of the sine function, the imaginary component also shows promise for bidirectional-attention-based diffusion language models (Nie et al., 2025; Ye et al., 2025), and we will provide experiments on these aspects in follow-up.

6 Conclusion

We introduce RoPE++, which employs both real and imaginary attentions. Mathematical analysis first reveals the imaginary attention's potential for modeling long-context dependencies. Building upon this, we re-incorporate the originally discarded imaginary attention as a new group of heads while preserving the unified absolute–relative position embedding format. Particularly, we introduce RoPE++$_{\text{EH}}$, with equal head as well as halved cache, and RoPE++$_{\text{EC}}$ with equal cache and doubled heads. Pre-training and evaluation at 376M and 776M model sizes show that both RoPE++$_{\text{EH}}$ and RoPE++$_{\text{EC}}$ outperform vanilla RoPE and other position embeddings on average across short-context tasks and acquire even larger gains in long-context scenarios. Further analysis confirms that imaginary attentions are more dominant in long-context modeling compared with original real attention, validating their effectiveness in enhancing long-context LLMs.

Ethical Statement

This research follows established ethical standards and practice principles. To our knowledge, our study processes no sensitive personal data, involves no human subjects, and targets no ethically risky applications. All experiments and analyses comply with recognized guidelines, ensuring integrity, transparency, and reliability.

Reproducibility Statement

To ensure the reproducibility of and to support the open-source community, we will publicly release RoPE++, its trained checkpoints, and the complete training and evaluation code. We expect these as a reference for future work on long-context LLMs, facilitating progress in this field.

References

Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. Gqa: Training generalized multi-query transformer models from multi-head checkpoints. *arXiv preprint arXiv:2305.13245*, 2023.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. PIQA: reasoning about physical commonsense in natural language. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pp. 7432–7439. AAAI Press, 2020. doi: 10.1609/AAAI.V34I05.6239. URL `https://doi.org/10.1609/aaai.v34i05.6239`.

bloc97. Dynamically scaled rope further increases performance of long context llama with zero fine-tuning, July 2023. URL `https://www.reddit.com/r/LocalLLaMA/comments/14mrgpr/dynamically_scaled_rope_further_increases/`.

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, Xiaoyi Dong, Haodong Duan, Qi Fan, Zhaoye Fei, Yang Gao, Jiaye Ge, Chenya Gu, Yuzhe Gu, Tao Gui, Aijia Guo, Qipeng Guo, Conghui He, Yingfan Hu, Ting Huang, Tao Jiang, Penglong Jiao, Zhenjiang Jin, Zhikai Lei, Jiaxing Li, Jingwen Li, Linyang Li, Shuaibin Li, Wei Li, Yining Li, Hongwei Liu, Jiangning Liu, Jiawei Hong, Kaiwen Liu, Kuikun Liu, Xiaoran Liu, Chengqi Lv, Haijun Lv, Kai Lv, Li Ma, Runyuan Ma, Zerun Ma, Wenchang Ning, Linke Ouyang, Jiantao Qiu, Yuan Qu, Fukai Shang, Yunfan Shao, Demin Song, Zifan Song, Zhihao Sui, Peng Sun, Yu Sun, Huanze Tang, Bin Wang, Guoteng Wang, Jiaqi Wang, Jiayu Wang, Rui Wang, Yudong Wang, Ziyi Wang, Xingjian Wei, Qizhen Weng, Fan Wu, Yingtong Xiong, Chao Xu, Ruiliang Xu, Hang Yan, Yirong Yan, Xiaogui Yang, Haochen Ye, Huaiyuan Ying, Jia Yu, Jing Yu, Yuhang Zang, Chuyu Zhang, Li Zhang, Pan Zhang, Peng Zhang, Ruijie Zhang, Shuo Zhang, Songyang Zhang, Wenjian Zhang, Wenwei Zhang, Xingcheng Zhang, Xinyue Zhang, Hui Zhao, Qian Zhao, Xiaomeng Zhao, Fengzhe Zhou, Zaida Zhou, Jingming Zhuo, Yicheng Zou, Xipeng Qiu, Yu Qiao, and Dahua Lin. Internlm2 technical report, 2024.

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.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the AI2 reasoning challenge. *CoRR*, abs/1803.05457, 2018. URL `http://arxiv.org/abs/1803.05457`.

OpenCompass Contributors. Opencompass: A universal evaluation platform for foundation models. `https://github.com/open-compass/opencompass`, 2023.

Chang Dai, Hongyu Shan, Mingyang Song, and Di Liang. Hope: Hyperbolic rotary positional encoding for stable long-range dependency modeling in large language models. *arXiv preprint arXiv:2509.05218*, 2025.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*, 2019.

Tri Dao. Flashattention-2: Faster attention with better parallelism and work partitioning. In *The Twelfth International Conference on Learning Representations*, 2024.

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

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

Demis Hassabis and Koray Kavukcuoglu. Introducing gemini 2.0: our new ai model for the agentic era, 2024. URL https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/.

Adi Haviv, Ori Ram, Ofir Press, Peter Izsak, and Omer Levy. Transformer language models without positional encodings still learn positional information. *arXiv preprint arXiv:2203.16634*, 2022.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. Ruler: What's the real context size of your long-context language models? *arXiv preprint arXiv:2404.06654*, 2024.

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. *arXiv preprint arXiv:2412.17739*, 2024.

HuggingFace. Open llm leaderboard. 2023. URL https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard.

InternLM. Internlm3-8b, January 2025. URL https://huggingface.co/internlm/internlm3-8b-instruct.

Hongye Jin, Xiaotian Han, Jingfeng Yang, Zhimeng Jiang, Zirui Liu, Chia-Yuan Chang, Huiyuan Chen, and Xia Hu. Llm maybe longlm: Self-extend llm context window without tuning. *arXiv preprint arXiv:2401.01325*, 2024.

Yuri Kuratov, Aydar Bulatov, Petr Anokhin, Ivan Rodkin, Dmitry Sorokin, Artyom Sorokin, and Mikhail Burtsev. Babilong: Testing the limits of llms with long context reasoning-in-a-haystack. *arXiv preprint arXiv:2406.10149*, 2024.

Yann LeCun, Yoshua Bengio, et al. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995, 1995.

ChiYan Lee, Hideyuki Hasegawa, and Shangce Gao. Complex-valued neural networks: A comprehensive survey. *IEEE/CAA Journal of Automatica Sinica*, 9(8):1406–1426, 2022.

Jeffrey Li, Alex Fang, Georgios Smyrnis, Maor Ivgi, Matt Jordan, Samir Yitzhak Gadre, Hritik Bansal, Etash Guha, Sedrick Scott Keh, Kushal Arora, et al. Datacomp-lm: In search of the next generation of training sets for language models. *Advances in Neural Information Processing Systems*, 37:14200–14282, 2024.

Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pp. 3214–3252. Association for Computational Linguistics, 2022. doi: 10.18653/V1/2022.ACL-LONG.229. URL https://doi.org/10.18653/v1/2022.acl-long.229.

Zhixuan Lin, Evgenii Nikishin, Xu Owen He, and Aaron Courville. Forgetting transformer: Softmax attention with a forget gate. *arXiv preprint arXiv:2503.02130*, 2025.

Hao Liu, Wilson Yan, Matei Zaharia, and Pieter Abbeel. World model on million-length video and language with ringattention. *arXiv e-prints*, pp. arXiv–2402, 2024a.

Xiaoran Liu, Ruixiao Li, Qipeng Guo, Zhigeng Liu, Yuerong Song, Kai Lv, Hang Yan, Linlin Li, Qun Liu, and Xipeng Qiu. Reattention: Training-free infinite context with finite attention scope. *arXiv preprint arXiv:2407.15176*, 2024b.

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*, 2024c.

Xiaoran Liu, Siyang He, Qiqi Wang, Ruixiao Li, Yuerong Song, Zhigeng Liu, Linlin Li, Qun Liu, Zengfeng Huang, Qipeng Guo, et al. Beyond homogeneous attention: Memory-efficient llms via fourier-approximated kv cache. *arXiv preprint arXiv:2506.11886*, 2025a.

Xiaoran Liu, Zhigeng Liu, Zengfeng Huang, Qipeng Guo, Ziwei He, and Xipeng Qiu. Longllada: Unlocking long context capabilities in diffusion llms. *arXiv preprint arXiv:2506.14429*, 2025b.

Ilya Loshchilov, Frank Hutter, et al. Fixing weight decay regularization in adam. *arXiv preprint arXiv:1711.05101*, 5(5):5, 2017.

Yi Lu, Xin Zhou, Wei He, Jun Zhao, Tao Ji, Tao Gui, Qi Zhang, and Xuanjing Huang. Longheads: Multi-head attention is secretly a long context processor. *arXiv preprint arXiv:2402.10685*, 2024.

Kai Lv, Xiaoran Liu, Qipeng Guo, Hang Yan, Conghui He, Xipeng Qiu, and Dahua Lin. Longwanjuan: Towards systematic measurement for long text quality. *arXiv preprint arXiv:2402.13583*, 2024.

Xin Men, Mingyu Xu, Bingning Wang, Qingyu Zhang, Hongyu Lin, Xianpei Han, and Weipeng Chen. Base of rope bounds context length. *arXiv preprint arXiv:2405.14591*, 2024.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL `https://openreview.net/forum?id=Byj72udxe`.

AI Meta. Introducing meta llama 3: The most capable openly available llm to date. *Meta AI.*, 2024a.

AI Meta. Llama 3.2: Revolutionizing edge ai and vision with open, customizable models. *Meta AI.*, 2024b.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? A new dataset for open book question answering. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pp. 2381–2391. Association for Computational Linguistics, 2018. doi: 10.18653/V1/D18-1260. URL `https://doi.org/10.18653/v1/d18-1260`.

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. Large language diffusion models. *arXiv preprint arXiv:2502.09992*, 2025.

OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

OpenAI. O1: Openai's first model, 2024. URL `https://openai.com/o1/`. Accessed: 2024-12-25.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Quan Ngoc Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. The LAMBADA dataset: Word prediction requiring a broad discourse context. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics, 2016. doi: 10.18653/V1/P16-1144. URL `https://doi.org/10.18653/v1/p16-1144`.

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.

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

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level google-proof q&a benchmark. *CoRR*, abs/2311.12022, 2023. doi: 10.48550/ARXIV.2311.12022. URL https://doi.org/10.48550/arXiv.2311.12022.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pp. 8732–8740. AAAI Press, 2020. doi: 10.1609/AAAI.V34I05.6399. URL https://doi.org/10.1609/aaai.v34i05.6399.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. Social iqa: Commonsense reasoning about social interactions. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pp. 4462–4472. Association for Computational Linguistics, 2019. doi: 10.18653/V1/D19-1454. URL https://doi.org/10.18653/v1/D19-1454.

Jürgen Schmidhuber, Sepp Hochreiter, et al. Long short-term memory. *Neural Comput*, 9(8): 1735–1780, 1997.

Jianlin Su. Rerope: Rectified rotary position embeddings, July 2023. URL https://github.com/bojone/rerope.

Jianlin Su. Transformer upgrade path: 17. insights into multimodal positional embedding, March 2024a. URL https://spaces.ac.cn/archives/10040.

Jianlin Su. Transformer upgrade path: 18. rope base selection principle, March 2024b. URL https://kexue.fm/archives/10122.

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

Tianxiang Sun, Xiaotian Zhang, Zhengfu He, Peng Li, Qinyuan Cheng, Xiangyang Liu, Hang Yan, Yunfan Shao, Qiong Tang, Shiduo Zhang, Xingjian Zhao, Ke Chen, Yining Zheng, Zhejian Zhou, Ruixiao Li, Jun Zhan, Yunhua Zhou, Linyang Li, Xiaogui Yang, Lingling Wu, Zhangyue Yin, Xuanjing Huang, Yu-Gang Jiang, and Xipeng Qiu. Moss: An open conversational large language model. *Machine Intelligence Research*, 2024. ISSN 2731-5398. doi: 10.1007/s11633-024-1502-8. URL https://github.com/OpenMOSS/MOSS.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. Superglue: A stickier benchmark for general-purpose language understanding systems. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 3261–3275, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/4496bf24afe7fab6f046bf4923da8de6-Abstract.html.

Feiyu Wang, Guoan Wang, Yihao Zhang, Shengfan Wang, Weitao Li, Bokai Huang, Shimao Chen, Zihan Jiang, Rui Xu, and Tong Yang. ifairy: the first 2-bit complex llm with all parameters in $\{\pm 1, \pm i\}$. *arXiv preprint arXiv:2508.05571*, 2025.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Xilin Wei, Xiaoran Liu, Yuhang Zang, Xiaoyi Dong, Pan Zhang, Yuhang Cao, Jian Tong, Haodong Duan, Qipeng Guo, Jiaqi Wang, et al. Videorope: What makes for good video rotary position embedding? *arXiv preprint arXiv:2502.05173*, 2025.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, et al. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pp. 38–45, 2020.

Chaojun Xiao, Pengle Zhang, Xu Han, Guangxuan Xiao, Yankai Lin, Zhengyan Zhang, Zhiyuan Liu, and Maosong Sun. Infllm: Training-free long-context extrapolation for llms with an efficient context memory. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Wenhan Xiong, Jingyu Liu, Igor Molybog, Hejia Zhang, Prajjwal Bhargava, Rui Hou, Louis Martin, Rashi Rungta, Karthik Abinav Sankararaman, Barlas Oguz, et al. Effective long-context scaling of foundation models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 4643–4663, 2024.

Hang Yan, Bocao Deng, Xiaonan Li, and Xipeng Qiu. Tener: adapting transformer encoder for named entity recognition. *arXiv preprint arXiv:1911.04474*, 2019.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a.

Songlin Yang, Yikang Shen, Kaiyue Wen, Shawn Tan, Mayank Mishra, Liliang Ren, Rameswar Panda, and Yoon Kim. Path attention: Position encoding via accumulating householder transformations. *arXiv preprint arXiv:2505.16381*, 2025b.

Jiacheng Ye, Zhihui Xie, Lin Zheng, Jiahui Gao, Zirui Wu, Xin Jiang, Zhenguo Li, and Lingpeng Kong. Dream 7b: Diffusion large language models. *arXiv preprint arXiv:2508.15487*, 2025.

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, et al. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*, 2024.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? In Anna Korhonen, David R. Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pp. 4791–4800. Association for Computational Linguistics, 2019. doi: 10.18653/V1/P19-1472. URL `https://doi.org/10.18653/v1/p19-1472`.

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. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024a.

Chuanyang Zheng, Yihang Gao, Han Shi, Jing Xiong, Jiankai Sun, Jingyao Li, Minbin Huang, Xiaozhe Ren, Michael Ng, Xin Jiang, et al. Dape v2: Process attention score as feature map for length extrapolation. *arXiv preprint arXiv:2410.04798*, 2024b.

A USE OF LARGE LANGUAGE MODELS

We use Large Language Models solely for language-centric assistance, including checking grammar, style, and clarity. No aspect of research, including ideation, experimental design, or scientific contribution, is influenced or generated by the output of LLMs.

B Preliminary: key properties of RoPE

Rotary Position Embedding (RoPE) encodes absolute positions by splitting the feature dimensions of query and key vectors $\boldsymbol{q}_t, \boldsymbol{k}_s$ into 2-D pairs and rotating each pair (Su et al., 2024). The rotation angle is the product of the token index t or s, and θ_n. Owing to the properties of rotation matrices, the independently applied absolute position embedding on $\boldsymbol{q}_t, \boldsymbol{k}_s$ fuse into a relative position embedding, namely $\cos\theta_n(t-s), \sin\theta_n(t-s)$, of the attention matrix, as shown in Equation 6.

$$\begin{aligned} \boldsymbol{A}_{t,s} &= \underbrace{\sum_{n=0}^{d/2-1} \begin{bmatrix} q_t^{(2n)} \\ q_t^{(2n+1)} \end{bmatrix}^\top \begin{bmatrix} \cos\theta_n(t-s) & \sin\theta_n(t-s) \\ -\sin\theta_n(t-s) & \cos\theta_n(t-s) \end{bmatrix} \begin{bmatrix} k_s^{(2n)} \\ k_s^{(2n+1)} \end{bmatrix}}_{\text{Relative PE}} \\ &= \underbrace{\sum_{n=0}^{d/2-1} \left(\begin{bmatrix} \cos\theta_n t & -\sin\theta_n t \\ \sin\theta_n t & \cos\theta_n t \end{bmatrix} \begin{bmatrix} q_t^{(2n)} \\ q_t^{(2n+1)} \end{bmatrix} \right)^\top \left(\begin{bmatrix} \cos\theta_n s & -\sin\theta_n s \\ \sin\theta_n s & \cos\theta_n s \end{bmatrix} \begin{bmatrix} k_s^{(2n)} \\ k_s^{(2n+1)} \end{bmatrix} \right)}_{\text{Absolute PE}} \end{aligned} \tag{6}$$

By default, the rotary angles $\theta_n = 10000^{-2n/d}, n = 0, \cdots, d/2-1$.

Equation 6 presents RoPE in vector form. Since any 2-D vector corresponds to a complex number, the rotation of such a vector is equivalent to complex multiplication.

$$\tilde{q}_t^{(n)} = q_t^{(2n)} + i \cdot q_t^{(2n+1)}, \quad \tilde{k}_s^{(n)} = k_s^{(2n)} + i \cdot k_s^{(2n+1)}$$

Building on this equivalence, RoPE can be expressed in complex form as shown in Equation 1.

Besides unifying relative and absolute position embeddings, RoPE exhibits semantic aggregation and long-context decay (Su et al., 2024). On one hand, when $\boldsymbol{q}, \boldsymbol{k}$ vectors are semantically close, their attention score remains large on average, regardless of relative distance Δt. This property is detailed in Su (2024b). If we have a vector $\boldsymbol{k}$ that is independent and identically distributed with respect to $\boldsymbol{q}$, with average μ and variance σ^2 for every feature dimension, and a vector that is only slightly perturbed with respect to $\boldsymbol{q} + \boldsymbol{\varepsilon}$, the expected attention score difference can be calculated as follows and proved to be positive.

$$\begin{aligned} &\mathbb{E}_{\boldsymbol{q},\boldsymbol{k},\boldsymbol{\varepsilon}} \left[\boldsymbol{q}^\top \boldsymbol{\mathcal{R}}_{-\Delta t} (\boldsymbol{q} + \boldsymbol{\varepsilon}) - \boldsymbol{q}^\top \boldsymbol{\mathcal{R}}_{-\Delta t} \boldsymbol{k} \right] \\ =&\mathbb{E}_{\boldsymbol{q}} \left[\boldsymbol{q}^\top \boldsymbol{\mathcal{R}}_{-\Delta t} \boldsymbol{q} \right] - \mathbb{E}_{\boldsymbol{q},\boldsymbol{k}} \left[\boldsymbol{q}^\top \boldsymbol{\mathcal{R}}_{-\Delta t} \boldsymbol{k} \right] \\ =&\mathbb{E}_{\boldsymbol{q}} \left[\boldsymbol{q}^\top \boldsymbol{\mathcal{R}}_{-\Delta t} \boldsymbol{q} \right] - \mathbb{E}_{\boldsymbol{q}}[\boldsymbol{q}]^\top \boldsymbol{\mathcal{R}}_{-\Delta t} \mathbb{E}_{\boldsymbol{k}}[\boldsymbol{k}] \\ =&\mathbb{E}_{\boldsymbol{q}} \left[\boldsymbol{q}^\top \boldsymbol{\mathcal{R}}_{-\Delta t} \boldsymbol{q} \right] - \mu^2 \boldsymbol{1}^\top \boldsymbol{\mathcal{R}}_{-\Delta t} \boldsymbol{1} \\ =&\mathbb{E}_{\boldsymbol{q}} \left[\sum_{n=0}^{d/2-1} \left({q^{(2n)}}^2 + {q^{(2n+1)}}^2 \right) \cos\left(-\theta_n \Delta t\right) \right] - \sum_{n=0}^{d/2-1} 2\mu^2 \cos\left(-\theta_n \Delta t\right) \\ =& \sum_{n=0}^{d/2-1} 2\left(\mu^2 + \sigma^2\right) \cos\theta_n \Delta t - \sum_{n=0}^{d/2-1} 2\mu^2 \cos\theta_n \Delta t \\ =& \sum_{n=0}^{d/2-1} 2\sigma^2 \cos\left(10000^{-\frac{2n}{d}} \Delta t\right) > 0 \end{aligned}$$

On the other hand, as Δt increases, the attention between any $\boldsymbol{q}, \boldsymbol{k}$ decreases on average. We can similarly derive this by showing that the expectation of the attention score, as shown below, is almost monotonically decaying with the increase of Δt.

$$\begin{aligned} \mathbb{E}_{\boldsymbol{q},\boldsymbol{k}} \left[\boldsymbol{q}^\top \boldsymbol{\mathcal{R}}_{-\Delta t} \boldsymbol{k} \right] &= \mathbb{E}_{\boldsymbol{q}} \left[\sum_{n=0}^{d/2-1} \left(q^{(2n)} k^{(2n)} + q^{(2n+1)} k^{(2n+1)} \right) \cos\left(\theta_n \Delta t\right) \right] \\ &= \sum_{n=0}^{d/2-1} 2(\mu^2 + \sigma^2) \cos\left(10000^{-\frac{2n}{d}} \Delta t\right) \end{aligned}$$

Both properties arise from averaging $\cos(\theta\Delta t)$ over frequency θ sampled based on $\theta_n = 10000^{-2n/d}, n = 0, \cdots, d/2-1$. It is a discrete approximation $c_{\text{Re}}(\Delta t)$ to a cosine integral function $\tilde{c}_{\text{Re}}(\Delta t)$, as shown in Equation 7. We refer to this as the characteristic curve of RoPE, as shown in Figure 1. It is positive and decaying, conferring these two mathematical properties of RoPE.

$$c_{\text{Re}}(\Delta t) = \frac{2}{d}\sum_{n=0}^{d/2-1} \cos\left(10^{-\frac{8n}{d}}\Delta t\right), \quad \tilde{c}_{\text{Re}}(\Delta t) = \int\limits_{10^{-4}}^{1} \frac{\cos\theta t}{\theta \ln 10^4} \mathrm{d}\theta = \mathrm{Ci}(\Delta t) - \mathrm{Ci}\left(\frac{\Delta t}{10^4}\right) \tag{7}$$

C More Experimental Details

The configuration of our 376M and 776M models can be summarized in the following table. Our models use the same tokenizer as the Llama 3 Series (Meta, 2024b;a; Dubey et al., 2024).

	376M	776M
Hidden Size	1024	1536
Intermediate Size	3584	5376
Num Layer	8	12
Num Attn Head	8	12
Num KV Head	4	6
Vocab Size	128256	128256

Table 3: The hyper-parameter of different model sizes.