

SCALING LAWS OF ROPE-BASED EXTRAPOLATION

Xiaoran Liu, Hang Yan, Shuo Zhang, Chenxin An, Xipeng Qiu, Dahua Lin

School of Computer Science, Fudan University
Shanghai AI Lab

ABSTRACT

The extrapolation capability of Large Language Models (LLMs) based on Rotary Position Embedding Su et al. (2021) is currently a topic of considerable interest. The mainstream approach to addressing extrapolation with LLMs involves modifying RoPE by replacing 10000, the rotary base of $\theta_n = 10000^{-2n/d}$ in the original RoPE, with a larger value and providing longer fine-tuning text. In this work, we first observe that fine-tuning a RoPE-based LLM with either a smaller or larger base in pre-training context length could significantly enhance its extrapolation performance. After that, we propose *Scaling Laws of RoPE-based Extrapolation*, a unified framework from the periodic perspective, to describe the relationship between the extrapolation performance and base value as well as tuning context length. In this process, we also explain the origin of the RoPE-based extrapolation issue by *critical dimension for extrapolation*. Besides these observations and analyses, we achieve extrapolation up to 1 million context length within only 16K training length on LLaMA2 7B and 13B (Touvron et al., 2023b).

1 INTRODUCTION

Large Language Models (LLMs) have become the dominant architecture in a variety of natural language processing tasks(OpenAI, 2023; Touvron et al., 2023a;b), while Transformers (Vaswani et al., 2017) based on Rotary Position Embedding (RoPE) (Su et al., 2021) have become the dominant backbone in wide range of LLM design (Chowdhery et al., 2022; Nijkamp et al., 2022; Touvron et al., 2023a;b). While RoPE can theoretically represent sequences through trigonometric functions, as detailed in Appendix A, its performance drops when the input sequence or context length surpasses the training length(Press et al., 2021; Chen et al., 2023), seen in Figure 1. This *extrapolation problem* (Press et al., 2021) limits tasks like long text modeling and summarization(An et al., 2023).

Concerning the extrapolation issue with RoPE, different works have provided various interpretations and corresponding solving attempts. These works could divided into two schools of thought. One limits the scope of self-attention (Ratner et al., 2022; Han et al., 2023) given the fact that self-attention computations in RoPE fail to keep stable beyond training context and exhibit attention

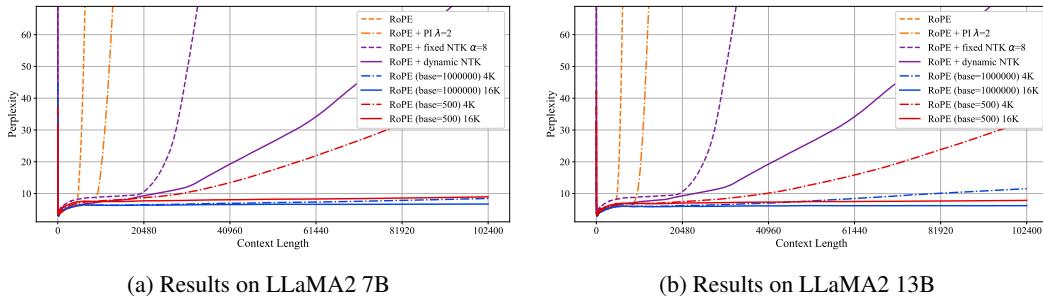


Figure 1: Perplexity on the validation data of Books3 (Presser, 2020) with different extrapolation methods, including Dynamic NTK (LocalLLaMA, 2023b). RoPE fine-tuned with either a smaller or larger base on the original training length of 4K or a much longer context of 16K, could outperform other extrapolation strategies and extrapolate to 100K context length.

score explosion as well as monotonous entropy increase (Chen et al., 2023; Han et al., 2023). The other aims to capture longer contexts by using smaller rotation angles and longer fine-tuning context (Chen et al., 2023; Peng et al., 2023). Currently, popular methods, such as Dynamic NTK (LocalLLaMA, 2023a) and Code LLaMA (Rozière et al., 2023), mainly come from the second approach. Both approaches adapt RoPE to longer contexts with a larger rotary base. Specifically, Dynamic NTK (LocalLLaMA, 2023a) adjusts the base with a coefficient increasing with the length of inference, allowing RoPE-based LLMs to adapt simultaneously to longer context, while Code LLaMA (Rozière et al., 2023) directly sets the base at 1000000 and gets further trained on sequences with 16K tokens, yielding a context beyond 100K.

While recent studies have shown promising results, they have primarily focused on specific base values and tuning context lengths. This leaves a notable gap in understanding how base value, tuning length, and extrapolation performance relate. For instance, while larger bases improve extrapolation in models like LLaMA2 (Touvron et al., 2023a), surprisingly, we also find that fine-tuning with smaller bases with the original training length is also conducive to the extrapolation capability of LLaMA2, which is also demonstrated in Figure 1. Furthermore, when trained in a longer context, RoPE with a smaller base can match or even surpass those with a larger one. At the same time, fine-tuning with a base of 1000000 on the original training length also achieves extrapolation up to 100K. These findings pose several questions. ***Q1: Is 10000 the worst base value for extrapolation in the fine-tuning phase? Q2: Is there a mathematical relationship between rotary base, training context length, and extrapolation limit? Q3: If so, can we achieve unbound extrapolation accordingly?***

In this paper, we conduct further experiments on increasing and decreasing the rotary base in Section 2 and subsequently discover that adjusting the rotary base in both directions can contribute to the extrapolation of RoPE-based LLMs. Building upon these observations, we provide a comprehensive explanation for the seemingly counter-intuitive phenomenon from a periodic perspective. Meanwhile, we establish a unified theoretical framework for RoPE-based extrapolation known as the *Scaling Laws of RoPE-based Extrapolation*¹. We pinpoint specific changes during base reduction that lead to a significant boost in extrapolation in Section 3.1 and identify the upper bound of extrapolation for larger bases in Section 3.3. This clarifies how Code LLaMA (Rozière et al., 2023) manages a 100K extrapolation with only 16K training context. Furthermore, we validate our theories in Section 3.4 and Appendix C, shedding light on both the core principles of Dynamic NTK (LocalLLaMA, 2023a) and pinpointing instability sources self-attention computations in RoPE-based extrapolation (Han et al., 2023). Finally, we present the contributions and guiding significance of this work for other methods that achieve extrapolation during the inference phase. In summary, our contribution can be summarized as follows:

- We first highlight a surprisingly strange phenomenon 10000 is the worst base value for RoPE-based extrapolation in the fine-tuning phase. Remarkably, fine-tuning with either a larger or smaller base within the training context length greatly enhances extrapolation, which provides a new vision to the extrapolation research of RoPE (Su et al., 2021).
- Then we introduce a unified theoretical framework for RoPE-based extrapolation from a periodic perspective, known as the *Scaling Laws of RoPE-based Extrapolation*, which not only clarifies the aforementioned observations and addresses unanswered questions in existing research (Rozière et al., 2023), but also discover the *Critical Dimension for RoPE-based Extrapolation*, revealing the underlying reasons for the extrapolation issue of RoPE.
- Finally, for extrapolation within a defined context, we present the suggested fine-tuning base value determined by the context limit and extend the context of LLaMA2 7B and 13B (Touvron et al., 2023b) to surpass 100K tokens by tuning RoPE with base 1000000 and 4K tuning length. For unpredictable extrapolation, we propose a RoPE with a smaller base, such as 500, and achieve an almost 1M token context with a mere 16K tuning length.

¹It is important to note that the scaling laws proposed in this work are *irrelevant* with the well-known scaling laws (Kaplan et al., 2020). In this paper, *scale* refers to the adjustment of RoPE’s rotary angles (LocalLLaMA, 2023a), rather than the change of the model size. Unlike the scaling laws (Kaplan et al., 2020) that are empirically derived, our scaling laws define mathematical relations between context window size and rotary base, supported by experiments.

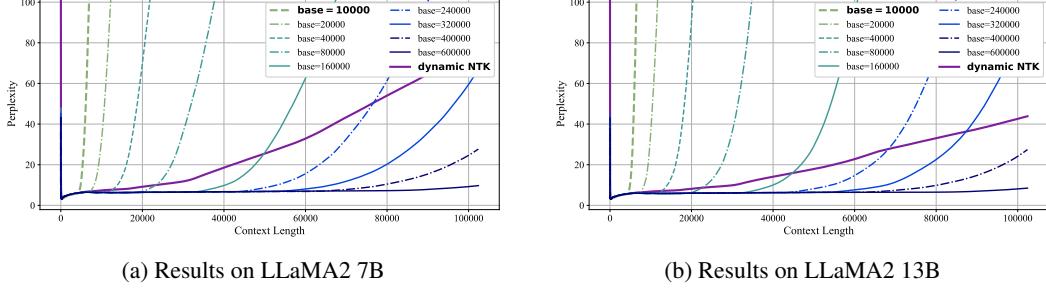


Figure 2: Perplexity of larger bases on Books3 (Presser, 2020) shows better extrapolation.

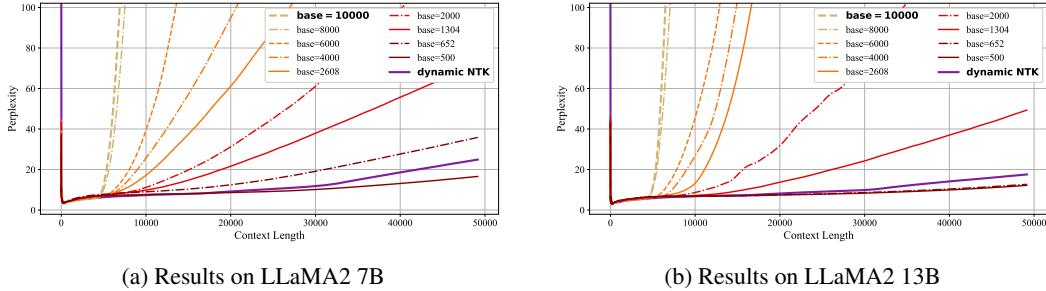


Figure 3: Perplexity of smaller bases on Books3 (Presser, 2020) shows better extrapolation.

2 OBSERVATION

2.1 LARGER BASES PROMISE BETTER EXTRAPOLATION

We first conduct the extrapolation experiments with larger bases, based on the experimental setup in Appendix B.1. It is evident that tuning with larger bases could significantly improve the extrapolation performance of RoPE as shown in Figure 2. Besides, there are several noteworthy points.

First, Larger bases allow LLaMA2 (Touvron et al., 2023b) to extrapolate beyond its training context length, aligning with findings from Rozière et al. (2023). Secondly, the extrapolation with larger bases has a clear limit where language modeling perplexity stays consistent. Beyond this limit, the extrapolation performance declines significantly. Furthermore, as the base value rises, LLaMA2 can extrapolate to a longer context. Finally, compared to Dynamic NTK (LocalLLaMA, 2023a), RoPE tuned with larger bases degrades much quicker beyond its extrapolation upper bound. Therefore, for fine-tuning with larger bases, the performance beyond the upper bound could be consistently overtaken by Dynamic NTK. Nevertheless, within the upper bound, this approach still outperforms Dynamic NTK by a considerable margin, leading to a context beyond 100K with only a 4K tuning length, when the base is set over 600000.

2.2 SMALLER BASES ALSO PROMISE BETTER EXTRAPOLATION

We then conduct the extrapolation experiments with smaller bases, using the same setup as for larger bases. Interestingly, even though this goes against common research findings (LocalLLaMA, 2023a; Rozière et al., 2023), fine-tuning RoPE with smaller bases on the original context length still boosts extrapolation, as shown in Figure 3. It also extends the context window beyond the training length. Yet, there are distinct differences when comparing RoPE with smaller bases to larger ones.

Firstly, RoPE with smaller bases does not have a distinct upper bound of extrapolation. While perplexity worsens as context length grows, this decline is gentler with smaller bases. Secondly, the enhancement in extrapolation of RoPE with smaller bases is not uniform. Between a base of 10000 and 8000, the extrapolation performance exhibits a tiny improvement. Then between a base of 8000 and 2608, improvement is moderate. After that, from 2608 to 1304 and further to 652, the improvement becomes more pronounced. Finally, when the rotary base is 500, the extrapolation

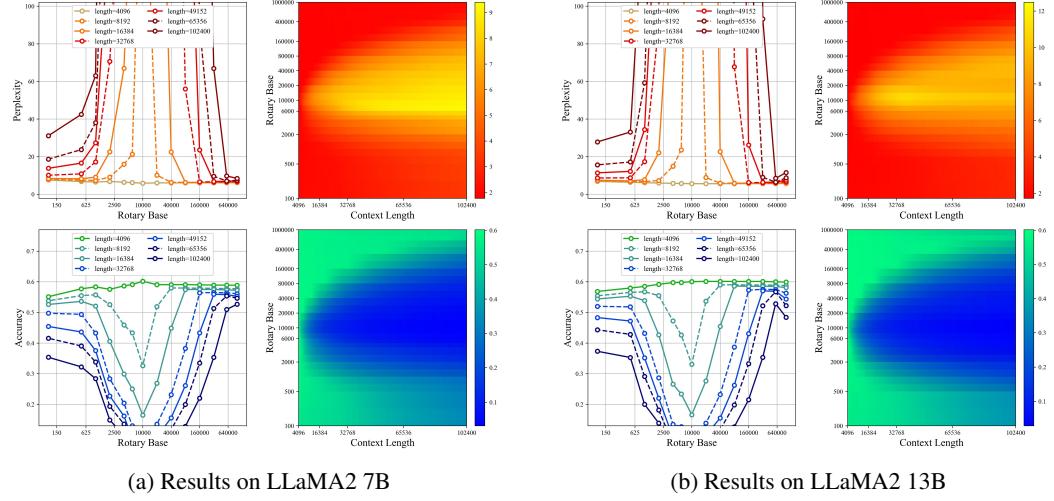


Figure 4: Perplexity and accuracy on the validation data from Books3 dataset (Presser, 2020) of LLaMA2 7B and 13B (Touvron et al., 2023b). In (a) and (b), the first row shows perplexity while the second row shows accuracy, both measured cumulatively. The first column shows the change of extrapolation performance w.r.t. rotary base at different context windows with a line plot, while the second column visualizes it with a heatmap. In the heatmap of perplexity, the value is log-scaled, namely, cross-entropy loss of language modeling.

curve becomes sufficiently smooth thus resulting in strong extrapolation over 48K context length and superior performance over Dynamic NTK (LocalLLaMA, 2023a).

Combining these two results, we observe a surprisingly strange phenomenon in RoPE-based extrapolation as depicted in Figure 4. Specifically, base 10000 yields the worst extrapolation performance when fine-tuned, thus answering Q1 in the Introduction. As the base either decreases or increases, performance notably improves. Interestingly, the improvements differ between the two directions. For larger bases, although the performance steadily improves, there exists a clear extrapolation upper bound. In contrast, for smaller bases, while the improvement is not uniform, the resulting extrapolation curve does not have an obvious breaking point. In the heatmap in Figure 4, there is a clear and continuous boundary for larger bases as well as a distinct transition phase for smaller bases.

3 EXPLANATION

3.1 SCALING LAW FOR SMALLER BASES

To understand how RoPE with smaller bases achieves impressive extrapolation within a constrained tuning context, it is crucial to explore the impact of reducing the base. As outlined in Appendix A, a smaller base amplifies the rotary angle $\theta_n = 10000^{-2n/d}$. This shortens T_n , the periods of $\sin(t-s)\theta_n$ or $\cos(t-s)\theta_n$ that RoPE uses to represent relative positions. Figure 5 shows the cosine waves for different dimensions to represent position information. In Figure 5a, it is evident that for smaller bases like 500, any period of $\cos(t-s)\theta_n$ is confined to 4096, i.e., the training length of LLaMA2 (Touvron et al., 2023b). In contrast, larger bases like 10000, extend the periods for several dimensions beyond the training length, as detailed in Section 3.2.

Hence, smaller bases lead to a broader range of cos or sin inputs during pre-training or fine-tuning, which ensures every dimension of q_t and k_s gets a well-trained representation. Additionally, as the base decreases, three pivotal points emerge, $\pi/2$, π , and 2π . Only when the cos or sin inputs in every dimension span from 0 to $\pi/2$ during training does the RoPE-based LLM recognize the negative values of cos and the non-monotonicity nature of sin. Similarly, the LLM becomes aware of the non-monotonicity of cos and negative sin values once inputs hit π . The RoPE-based LLM fully grasps the entire range of cos and sin only when the inputs surpass 2π , potentially embracing the periodicity of position embedding in every dimension. Then we propose Theorem 1. as follows.

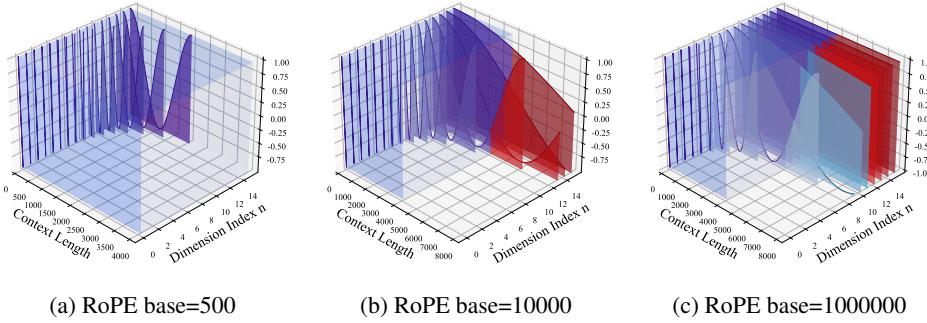


Figure 5: The visualized relationship among the period, training Length, and extrapolation. Consider a RoPE-based LLM with a head dimension size of 32, namely 16 rotary angles θ_n across various dimensions. Within each illustration, we visually represent the periods of $\cos(t-s)\theta_n$ for these dimensions using parallel purple planes. These are contrasted against the 4096 training context length, shown as a deep blue box. **(a)** For RoPE tuned with base 500, all periods of $\cos(t-s)\theta_n$ are limited within the training context. **(b)** For vanilla RoPE with base 10000, the periods of dimensions past the critical dimension (in red) stretch beyond the training context. **(c)** For RoPE tuned with base 1000000, although some updated periods colored (in sky blue) might surpass the training context, only dimensions past the critical dimension fail to extrapolate.

Theorem 1. (Scaling Law of Smaller Bases) For RoPE-based LLMs pre-trained with context length T_{train} , if we adjust the base to $\beta < 10000$ and conduct fine-tuning still with context length T_{train} , the extrapolation performance of RoPE-based LLMs will get improved. If β is reduced to $\beta_1, \beta_2, \beta_3$ as calculated below, the cos or sin inputs in every dimension will span from 0 to $\pi/2, \pi, 2\pi$ respectively, resulting in a more significant improvement.

$$\beta_1 = \frac{2T_{\text{train}}}{\pi}, \quad \beta_2 = \frac{T_{\text{train}}}{\pi}, \quad \beta_3 = \frac{T_{\text{train}}}{2\pi} \quad (1)$$

Particularly, for LLaMA2(Touvron et al., 2023b), where the context length T_{train} is 4096, we have $\beta_1 = 2608, \beta_2 = 1304, \beta_3 = 652$. It is worth noting that these three bases align with the pivotal points where improvement speeds up, as previously discussed during base reduction. Theorem 1 tells the improving paradigm of base reduction. Since it does not set an explicit extrapolation upper bound, RoPE with much smaller bases can potentially realize extrapolation to infinite context.

3.2 CRITICAL DIMENSION FOR EXTRAPOLATION

In comparison to much smaller bases like 500, where each period of $\cos(t-s)\theta_n$ fits within the training context, the default base in RoPE (Su et al., 2021), which is 10000, causes periods of certain dimensions to extend beyond the training context, as visualized in Figure 5b. Therefore, for RoPE-based LLMs, there exists a specific feature dimension, d_{extra} . For dimensions before d_{extra} , the periods of corresponding θ_n remain shorter than T_{train} , while for those after d_{extra} , the periods stretch beyond T_{train} . In other words, essentially, d_{extra} is the number of dimensions where $\cos(t-s)\theta_n$ and $\sin(t-s)\theta_n$ can cycle through their values within one period during pre-training or fine-tuning.

Consequently, for dimensions beyond d_{extra} , when RoPE-based LLMs extrapolate beyond T_{train} , the absolute position information of newly added tokens and the relative positional information in relation to previous tokens become out-of-distribution (OOD). This misalignment means the attention scores related to these dimensions, as illustrated in Equation 2, deviate from their expected distribution, causing a noticeable out-of-distribution in overall attention scores, thus leading to the extrapolation issue. We refer to this key dimension as the *Critical Dimension for RoPE-based*

extrapolation, which is formally defined and calculated as shown in Lemma 1.

$$\begin{aligned} \mathbf{A}_{t,s} &= \operatorname{Re} \left[\underbrace{\sum_{n=0}^{d/2-1} \tilde{q}_t^{(n)} \tilde{k}_s^{(n)*} e^{i(t-s)\theta_n}}_{\text{full attention scores in RoPE}} \right] \\ &= \operatorname{Re} \left[\underbrace{\sum_{n=0}^{d_{\text{extra}}/2-1} \tilde{q}_t^{(n)} \tilde{k}_s^{(n)*} e^{i(t-s)\theta_n}}_{\text{reliable part for extrapolation}} + \underbrace{\sum_{n=d_{\text{extra}}/2}^{d/2-1} \tilde{q}_t^{(n)} \tilde{k}_s^{(n)*} e^{i(t-s)\theta_n}}_{\text{OOD part for extrapolation}} \right]. \end{aligned} \quad (2)$$

Lemma 1. (Definition of Critical Dimension) For RoPE-based LLMs pre-trained with context length T_{train} , assuming that the size of self-attention head is d , there are at most the preceding d_{extra} dimensions that perceive complete periodic information thus receiving sufficient training for extrapolation, which is formally described as follows:

$$\begin{aligned} T_n &= \frac{2\pi}{\theta_n} = 2\pi \cdot 10000^{\frac{2n}{d}} \leq T_{\text{train}}, \quad \text{for } n = 0, \dots, d_{\text{extra}}/2 - 1, \\ T_n &= \frac{2\pi}{\theta_n} = 2\pi \cdot 10000^{\frac{2n}{d}} > T_{\text{train}}, \quad \text{for } n = d_{\text{extra}}/2, \dots, d/2 - 1. \end{aligned} \quad (3)$$

Then we define d_{extra} as the critical dimension for RoPE-based extrapolation and calculate it given

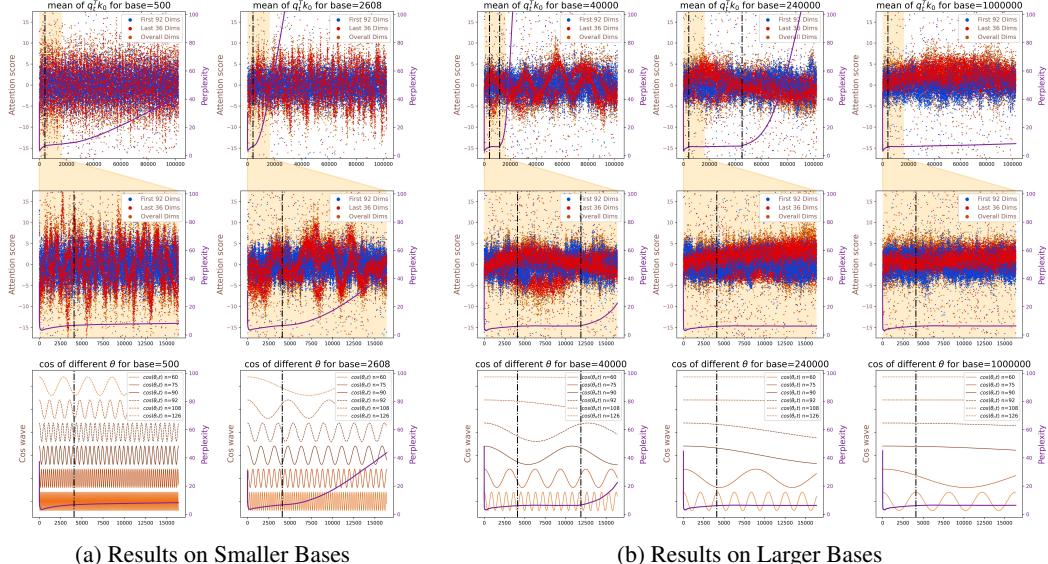
$$d_{\text{extra}} = 2 \left\lceil \frac{d}{2} \log_{10000} \frac{T_{\text{train}}}{2\pi} \right\rceil. \quad (4)$$

For LLaMA2(Touvron et al., 2023b), the critical dimension d_{extra} is 92. This implies that only the first 92 dimensions of the $\mathbf{q}_t, \mathbf{k}_s$ vectors of LLaMA2 have seen the complete positional information during the pre-training phase and are adequately trained. In other words, the last 36 dimensions lack sufficient training, contributing to the extrapolation challenges seen in RoPE-based LLMs (Chen et al., 2023; Han et al., 2023). The critical dimension plays a key role in enhancing extrapolation. A further discussion of the critical dimension is presented in Section 3.4. Here, we examine the attention score changes in the initial 92 versus the final 36 dimensions in relation to relative positions when the base is reduced. As is shown in Figure 6a, the attention scores of RoPE with smaller bases have effectively captured the oscillations from sin and cos within training, mitigating OOD concerns during extrapolation. Moreover, as the base becomes smaller, the perception becomes more comprehensive, resulting in improved extrapolation performance.

3.3 SCALING LAW FOR LARGER BASES

Based on the concept of the critical dimension, we can clarify the extrapolation results when fine-tuning RoPE with larger bases at the original context length. For LLaMA2 (Touvron et al., 2023b), since the periods of the first 92 dimensions fit within the training length, these feature dimensions start with a strong foundation for fine-tuning, adjusting to the new periodic shifts of positional embedding for extended contexts. Therefore, when RoPE is fine-tuned with a larger base value like 1000000, even though the tuning length is shorter than the extended periods corresponding to larger bases, these dimensions can still represent positional information correctly as is shown in Figure 5c.

However, for the last 36 dimensions, the absence of a full understanding of periodicity leads to overfitting. Furthermore, when the base is expanded, extending the period, these dimensions still fail to capture the entire positional information within the context length. So these dimensions are reliable only when the value of $\theta_n(t - s)$ is previously observed. Therefore, we can use the updated period of the critical dimension as an upper bound for extrapolation in RoPE-based LLM. As a result, we obtain Theorem 2., the scaling law for RoPE-based extrapolation with larger bases, which tells the relation between base value and extrapolation upper bound and thus answers Q2 in the Introduction.



(a) Results on Smaller Bases

(b) Results on Larger Bases

Figure 6: The relation between attention scores in first 92 and last 36 dimensions with the extrapolation performance in LLaMA 7B (Touvron et al., 2023b) fine-tuned with larger or smaller bases. The first row shows how average attention score in the first 92 and last 36 dimensions and the perplexity changes in 100K context length. The second row highlights the changes in the first 16K tokens. The third row visualizes that the period of critical dimension determines the attention score explosion and extrapolation issue. The black lines stand for training length or max context size.

Theorem 2. (Scaling Law of Larger Bases) For RoPE-based LLMs pre-trained with context length T_{train} , if we adjust the base to $\beta > 10000$ and conduct fine-tuning still with context length T_{train} , the extrapolation performance of RoPE-based LLMs will get improved. The extrapolation upper bound of RoPE-based LLM with larger bases, T_{extra} , is calculated as follows:

$$T_{\text{extra}} = 2\pi \cdot \beta^{d_{\text{extra}} \cdot \frac{1}{d}} = 2\pi \cdot \beta^{\lceil \frac{d}{2} \log_{10000} \frac{T_{\text{train}}}{2\pi} \rceil \cdot \frac{2}{d}}. \quad (5)$$

Inversely, if there is an expected extrapolation upper bound \tilde{T}_{extra} , then the smallest capable base β_0 is calculated as follows.

$$\beta_0 = 10000^{\log \frac{T_{\text{train}}}{2\pi} \frac{\tilde{T}_{\text{extra}}}{2\pi}}. \quad (6)$$

For extrapolation within a limited context, we can derive the suggested β_0 based on the expected context length. β_0 in Equation 6 is referred to as **critical base** for extrapolation and discussed in detail in Appendix C. To support the claims of Theorem 2, as illustrated in Figure 6b, we examined the attention scores of the first 92 and final 36 dimensions under various larger bases. Notably, while the attention scores of the first 92 dimensions remain relatively stable regardless of relative positions, the last 36 dimensions show significant variations. Reviewing the perplexity increase curves, it is evident that once the context length surpasses the extrapolation upper bound, the last 36 dimensions encounter unfamiliar positional information, leading to OOD attention scores and a sharp rise in perplexity. To further validate Theorem 2, we compare the max supported context lengths for different bases with the extrapolation upper bound derived from Theorem 2 in Figure 7. Impressively, there is a remarkable alignment between empirical results and theoretical predictions.

3.4 FURTHER VALIDATION FOR EXTRAPOLATION

To further illustrate the causal relationship between critical dimension and extrapolation, we undertake the following three experiments. First, we set a max index of 4096 for the position embedding of the last 36 dimensions of q_t, k_s in LLaMA2 7B (Touvron et al., 2023b). We find that apart from an outlier in one attention head, the attention score variations are considerably reduced, leading to better extrapolation. Second, we visualize the attention score of the first 92 and the last 36 dimensions

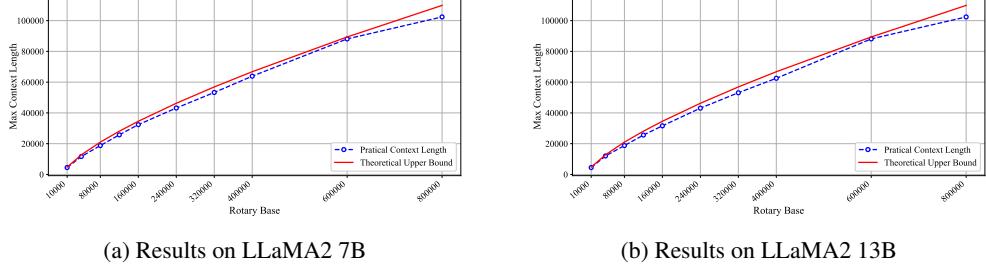


Figure 7: The comparison between max practical context length and extrapolation upper bound predicted by Theorem 2. for RoPE tuned with larger bases on pre-training context. For bases larger than 800000, the extrapolation upper bound is beyond the evaluating context length.

during extrapolation with Dynamic NTK (LocalLLaMA, 2023a) in Figure 8a. Consistent with our theory, the attention scores of the last 36 dimensions display enhanced consistency compared with straightforward extrapolation. Lastly, we remove the final 36 dimensions of q_t, k_s in LLaMA2 7B (Touvron et al., 2023b) and fine-tuned using the setup in Appendix B.1. Remarkably, post-trimming fine-tuning substantially surpassed direct fine-tuning, allowing extrapolation beyond 20K tokens, as illustrated in Figure 8b. This offers compelling evidence for the correlation between the critical dimension, attention score shifts, and extrapolation upper bound. It confirms that interpreting and enhancing RoPE-based LLM extrapolation from a periodic viewpoint is both valid and effective.

In summary, we establish a comprehensive framework from a periodic perspective and identify the fundamental factors, the critical dimension for RoPE-based extrapolation. This not only unveils the root causes of attention score explosion highlighted in Chen et al. (2023) and Han et al. (2023), but also intuitively demonstrates how adjusting the base during fine-tuning can elevate the extrapolation capabilities of RoPE-based LLMs. Furthermore, we answer Q3 in the Introduction and introduce the **Extended Scaling Law of RoPE-based Extrapolation** in Appendix C, the combination of Theorem 1., Lemma 1. and Theorem 2., for tuning in a longer context. Additionally, we also present the instructive value of our theory for other works focused on achieving longer context in the testing phase in Appendix D. Finally, as depicted in Figure 1, tuning RoPE with a base of 500 or 1000000 can

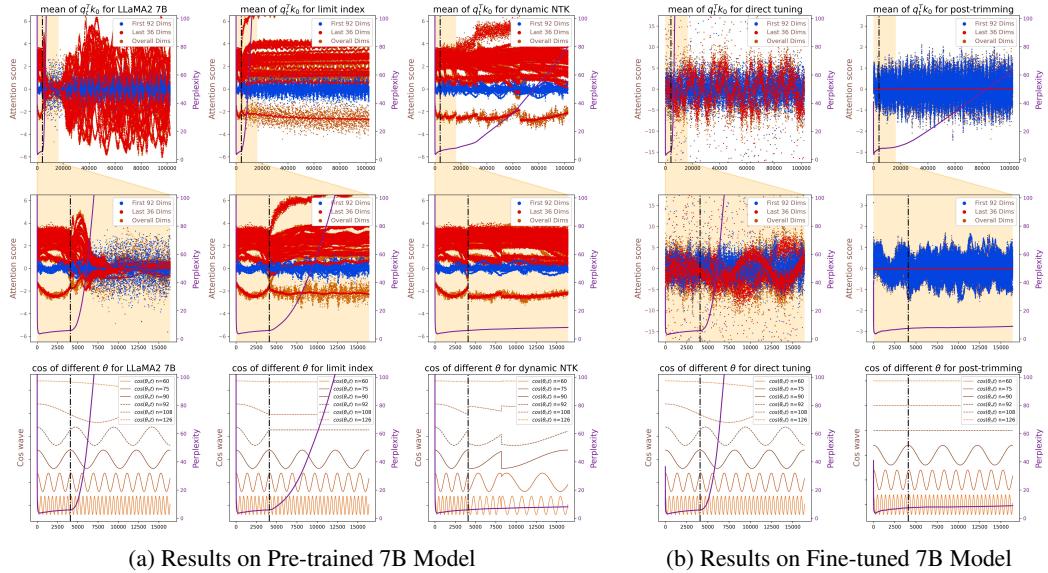


Figure 8: The relation between attention scores in first 92 and last 36 dimensions with the extrapolation performance in LLaMA 7B (Touvron et al., 2023b) evaluated or fine-tuned with original bases 10000. The meaning of each row is the same as that in Figure 6.

	128K	256K	512K	1M
base=500	9.49	12.41	23.99	51.28
base=500 log-scaled	9.13	10.01	12.07	19.07
base=1000000	7.07	76.82	1755.87	4914.81

Table 1: Perplexity on Books3 of LLaMA2 7B with base 500 and 1000000 tuned on 16K context.

both outperform Linear PI (Chen et al., 2023) and NTK method (LocalLLaMA, 2023b;a). Besides, we compare the performance of LLaMA2 7B Touvron et al. (2023b) with base 500 and 1000000 in 1M context length in Table 1 and find that for extrapolating to unpredictable length, RoPE with base 500, has remarkable advantages, especially when combined with log-scaled attention (Su, 2023b).

4 RELATED WORK

Recently, there have been a great deal of efforts devoted to expanding the context length of RoPE-based LLMs to 16K, 32K, or even 100K (Chen et al., 2023; LocalLLaMA, 2023b; Pal et al., 2023). Considering that LLMs, such as LLaMA2 (Touvron et al., 2023b), have already acquired sufficient knowledge in the pre-training stage and demonstrated excellence in short-context tasks, the emerging extrapolation improvements have primarily focused on the fine-tuning and the testing phase.

Extrapolating RoPE in Fine-tuning In the fine-tuning phase, current strategies primarily concentrate on using smaller rotary angles to depict long context features. For example, Chen et al. (2023) extends the context window to 16K by linearly reducing the rotary angles with $\lambda = T_{\text{extra}}/T_{\text{train}}$ to align the input position index within the original context size. Rozière et al. (2023) achieve 100K context window for code by increasing the rotary base to 1000000 and further training with 16K context length. Besides, Pal et al. (2023) also extends the context window to 32K by a new truncation strategy for refining the down-sampling method, which even sets some rotary angles to zero. Additionally, Peng et al. (2023) proposes a scaling function, fine-tunes LLaMA2 with 64K context, and finally achieves the 128K context length. Effective as they are, these methods all need additional tuning in longer contexts or face a disastrous collapse after the extrapolation bound.

Extrapolating RoPE in Inference In the inference phase, context window division or restriction is a well-discussed strategy for any LLM without further training. Effective as it is, it prohibits tokens of their deserved global attention (Han et al., 2023) or destroys the order between token chunks (Ratner et al., 2022; Su, 2023a), which is unsupported for long document summarization (Su, 2023a). Besides, it is worth noting the remarkable effect of the Neural Tangents Kernel (NTK) method (LocalLLaMA, 2023b), especially the dynamic version (LocalLLaMA, 2023a) as follows:

$$\text{for } \mathbf{A}_{t,s}, \theta_n = (10000 \cdot \alpha_t)^{-2n/d}, \text{ where } \alpha_t = \max \left(1, 2^{\lceil \log_2 \frac{t}{T_{\text{train}}} \rceil + 1} - 1 \right). \quad (7)$$

Dynamic NTK ² simply decreases the rotary angle exponentially with scaling coefficient α as a function w.r.t. the inference length, which is discussed in detail in Appendix D. Dynamic NTK enables LLaMA2(Touvron et al., 2023b) to extrapolate without any further tuning or window assisting and keeps good performance within 16K context. However, Dynamic NTK still presents an obvious degradation in performance around 32K context size as shown in Figure 1.

5 CONCLUSION

In summary, we initially highlight an intriguing observation: fine-tuning RoPE (Su et al., 2021) with either a larger or smaller base using the original pre-training context can boost the length extrapolation of RoPE-based LLMs. We then elucidate this observation through a unified theoretical lens rooted in a periodic view, *Scaling Laws of RoPE-based Extrapolation*. This framework clarifies the score of the RoPE-based extrapolation challenge and offers insights into how base modifications in fine-tuning or inference can enhance RoPE-based extrapolation. Finally, we provide the strategy for both extrapolating LLaMA2 (Touvron et al., 2023b) to a limited context and unpredictable inputs.

²In this work, we follow the form of scaling function in QWenLM (Alibaba, 2023)

REFERENCES

- Alibaba. Qwen technical report. Technical report, 2023. URL https://qianwen-res.oss-cn-beijing.aliyuncs.com/QWEN_TECHNICAL_REPORT.pdf.
- Chenxin An, Shansan Gong, Ming Zhong, Mukai Li, Jun Zhang, Lingpeng Kong, and Xipeng Qiu. L-eval: Instituting standardized evaluation for long context language models. *arXiv preprint arXiv:2307.11088*, 2023.
- 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.
- David Chiang and Peter Cholak. Overcoming a theoretical limitation of self-attention. *arXiv preprint arXiv:2202.12172*, 2022.
- Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.
- Tri Dao. Flashattention-2: Faster attention with better parallelism and work partitioning. *arXiv preprint arXiv:2307.08691*, 2023.
- Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.
- Chi Han, Qifan Wang, Wenhan Xiong, Yu Chen, Heng Ji, and Sinong Wang. Lm-infinite: Simple on-the-fly length generalization for large language models. *arXiv preprint arXiv:2308.16137*, 2023.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.
- Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*, 2017.
- Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.
- Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 2019.
- LocalLLaMA. Dynamically scaled rope further increases performance of long context llama with zero fine-tuning, July 2023a. URL https://www.reddit.com/r/LocalLLaMA/comments/14mrgpr/dynamically_scaled_rope_further_increases/.
- LocalLLaMA. Ntk-aware scaled rope allows llama models to have extended (8k+) context size without any fine-tuning and minimal perplexity degradation., June 2023b. URL https://www.reddit.com/r/LocalLLaMA/comments/14lz7j5/ntkaware_scaled_rope_allows_llama_models_to_have/.
- Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.
- Erik Nijkamp, Bo Pang, Hiroaki Hayashi, Lifu Tu, Huan Wang, Yingbo Zhou, Silvio Savarese, and Caiming Xiong. Codegen: An open large language model for code with multi-turn program synthesis. *arXiv preprint arXiv:2203.13474*, 2022.
- OpenAI. Gpt-4 technical report. Technical report, 2023.

- Arka Pal, Deep Karkhanis, Manley Roberts, Samuel Dooley, Arvind Sundararajan, and Siddartha Naidu. Giraffe: Adventures in expanding context lengths in llms. *arXiv preprint arXiv:2308.10882*, 2023.
- Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. Yarn: Efficient context window extension of large language models. *arXiv preprint arXiv:2309.00071*, 2023.
- Ofir Press, Noah A Smith, and Mike Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. *arXiv preprint arXiv:2108.12409*, 2021.
- Shawn Presser. Books3, 2020. URL <https://twitter.com/theshawwn/status/1320282149329784833>.
- Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. Zero: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, pp. 1–16. IEEE, 2020.
- Nir Ratner, Yoav Levine, Yonatan Belinkov, Ori Ram, Omri Abend, Ehud Karpas, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. Parallel context windows improve in-context learning of large language models. *arXiv preprint arXiv:2212.10947*, 2022.
- Baptiste Rozière, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Tal Remez, Jérémie Rapin, et al. Code llama: Open foundation models for code. *arXiv preprint arXiv:2308.12950*, 2023.
- Jianlin Su. Nbce: Naive bayes-based context extension, May 2023a.
- Jianlin Su. Improving transformer: Length extrapolation ability and position robustness. <https://spaces.ac.cn/archives/9444>, 2023b.
- Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *arXiv preprint arXiv:2104.09864*, 2021.
- Yutao Sun, Li Dong, Barun Patra, Shuming Ma, Shaohan Huang, Alon Benhaim, Vishrav Chaudhary, Xia Song, and Furu Wei. A length-extrapolatable transformer. *arXiv preprint arXiv:2212.10554*, 2022.
- 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*, 2023a.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Niko lay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023b.
- 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.
- Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32, 2019.

A PRELIMINARY

A.1 ROPE FROM SEQUENCE DOMAIN

Transformer models require the integration of explicit positional information through positional embeddings to effectively discern the order of input sequences (Vaswani et al., 2017). In this work, we direct our attention to the specific instance of positional encoding known as Rotary Position Embedding (RoPE) (Su et al., 2021), as prominently featured in the architecture of the LLaMA

model (Touvron et al., 2023a;b). Given a query vector $\mathbf{q}_t = [q_t^{(0)}, \dots, q_t^{(d-1)}] \in \mathbb{R}^d$ at position t and a key vector $\mathbf{k}_s = [k_s^{(0)}, \dots, k_s^{(d-1)}] \in \mathbb{R}^d$ at position s , RoPE first splits q_t, k_s into pairs on the direction of feature dimensions, with every two dimensions forming a complex number, or a vector in the complex plane as follows:

$$\begin{aligned}\tilde{\mathbf{q}}_t &= [\tilde{q}_t^{(0)}, \dots, \tilde{q}_t^{(d/2-1)}] & \tilde{q}_t^{(n)} &= q_t^{(2n)} + iq_t^{(2n+1)} \\ \tilde{\mathbf{k}}_s &= [\tilde{k}_s^{(0)}, \dots, \tilde{k}_s^{(d/2-1)}] & \tilde{k}_s^{(n)} &= k_s^{(2n)} + ik_s^{(2n+1)}.\end{aligned}\quad (8)$$

After that, RoPE injects the position information by an element-wise multiplication between the pre-processed $\tilde{\mathbf{q}}_t, \tilde{\mathbf{k}}_s$ and a list of θ -parameterized rotary vectors in the complex plane. When attention is calculated, relative position information $t - s$ is acquired through cos and sin.

$$\begin{aligned}\mathbf{A}_{t,s} &= \operatorname{Re} \left[(\tilde{\mathbf{q}}_t \odot e^{it\theta}) \cdot (\tilde{\mathbf{k}}_s \odot e^{is\theta})^T \right] \\ &= \operatorname{Re} \left[\sum_{n=0}^{d/2-1} \tilde{q}_t^{(n)} e^{it\theta_n} \left(\tilde{k}_s^{(n)} e^{is\theta_n} \right)^* \right] = \operatorname{Re} \left[\sum_{n=0}^{d/2-1} \tilde{q}_t^{(n)} \tilde{k}_s^{(n)*} e^{i(t-s)\theta_n} \right]. \\ &= \sum_{n=0}^{d/2-1} \left(q_t^{(2n)} k_s^{(2n)} + q_t^{(2n+1)} k_s^{(2n+1)} \right) \cos(t-s)\theta_n + \\ &\quad \sum_{n=0}^{d/2-1} \left(q_t^{(2n)} k_s^{(2n+1)} - q_t^{(2n+1)} k_s^{(2n)} \right) \sin(t-s)\theta_n\end{aligned}\quad (9)$$

While RoPE can theoretically convey the relative information at any context length, RoPE still fails to extrapolate practically. It is worth noting that rotary angles θ in Equation 9 play an important role. In the vanilla design of RoPE, θ is defined as Equation 10. Different angles correspond to different features and that is the starting point of most RoPE-based extrapolation methods (LocalLLaMA, 2023b; Rozière et al., 2023; Pal et al., 2023) shown in Table 2.

$$\theta = [\theta_0, \dots, \theta_{d/2-1}] \quad \theta_n = 10000^{-2n/d}. \quad (10)$$

A.2 ROPE FROM FREQUENCY DOMAIN

From a frequency domain perspective, the rotation operation of RoPE can be viewed as the Inverse Fourier Transform from the frequency domain to the time domain.

$$\begin{aligned}\mathbf{A}_{t,s} &= \operatorname{Re} \left[\sum_{n=0}^{d/2-1} \tilde{q}_t^{(n)} \tilde{k}_s^{(n)*} e^{i(t-s)\theta_n} \right]. \\ &= \operatorname{Re} \left[\mathcal{F}_{\theta}^{-1} \left[\tilde{\mathbf{q}}_t^{(n)} \tilde{\mathbf{k}}_s^{(n)*} \right] \right]\end{aligned}\quad (11)$$

At this juncture, the learning objective for RoPE-based LLM is essentially to understand features in terms of frequency. Depending on the value of θ_n , the higher dimensions correspond to the longer period as shown in Equation 12 as well as the lower-frequency features reflecting the longer contextual semantic relationship (Chen et al., 2023; Rozière et al., 2023).

$$T_n = \frac{2\pi}{\theta_n} = 2\pi \cdot 10000^{2n/d}, \quad \text{for } n = 0 \dots \frac{d}{2} - 1. \quad (12)$$

However, as illustrated in Figure 5, the trigonometric functions of lower-frequency features do not complete a full period within the training context. As a result, RoPE-based LLMs might not fully recognize the periodic nature of sin and cos waves, leading to inadequate training. Consequently, these lower-frequency features are more susceptible to under-fitting or over-fitting. Hence, the number of well-trained dimensions is essential and this is the critical dimension raised in Section 3.2.

Name	Applying Phase	Training Length	Context Window	Ratio	Auxiliary Window	Note
xPos (Sun et al., 2022)	Pre-training	1K	4K	×4	✓	BCA
Linear PI (Chen et al., 2023)	Fine-tuning	16K	16K	×1	✗	
Giraffe (Pal et al., 2023)	Fine-tuning	4K	32K	×8	✗	truncated
Code LLaMA (Rozière et al., 2023)	Fine-tuning	16K	100K+	×6	✗	
YaRN (Peng et al., 2023)	Fine-tuning	64K	128K	×2	✓	window ppl
Fixed NTK (LocalLLaMA, 2023b)	Inference	4K	16K	×4	✗	
Dynamic NTK (LocalLLaMA, 2023a)	Inference	4K	16K+	×4	✗	
PCW (Ratner et al., 2022)	Inference	2K	6K	×3	✓	for ICL
NBCE (Su, 2023a)	Inference	2K	10K	×5	✓	for ICL
LM-Infinite (Han et al., 2023)	Inference	4K	32K+	×8	✓	Λ-shaped
Scaling RoPE (Ours)	Fine-tuning	4K 16K	100K+ 1M+	×25 ×64	✗	base=1000000 base=500

Table 2: RoPE-based Extrapolation Strategies. For those applied in fine-tuning, the max training length is the tuning context length. For PCW and NBCE, the context size is the max training length times the optimal chunk number. It is worth noting that Scaling RoPE, namely tuning RoPE with smaller or larger bases discussed in this work, realizes the max context window and extension ratio.

B APPENDIX

B.1 EXPERIMENT SETUP

We conduct experiments on the pre-trained 7B and 13B LLaMA2 models (Touvron et al., 2023b). For fine-tuning 7B and 13B models, we use 32 A100 GPUs and adopt ZeRO3 strategies (Rajbhandari et al., 2020). We use AdamW (Loshchilov & Hutter, 2017) with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. We set the learning rate to 2×10^{-5} with no warmup. We set the max gradient norm to 2.5 for 7B and 1 for 13B respectively. We set the weight decay to zero.

For fine-tuning RoPE with different bases, we set the global batch size to 128, tuning the context length to 4K, the same as the training length, and the evaluating context length to 100K. We fine-tune the models for 1K steps using the next token prediction objective with training data from the Pile (Gao et al., 2020) and compare the tuning performance on the validation set of Books3 subset (Presser, 2020) from the Pile. We fine-tune LLaMA2 with CoLLiE³, a collaborative toolbox for tuning large language models in an efficient way, and conduct evaluation discussed in Appendix B.2 with OpenCompass⁴. Both training and testing are accelerated by FlashAttention-2 (Dao, 2023).

We compare results with mainstream extrapolating strategies, such as Linear Position Interpolation (Chen et al., 2023) and NTK method (LocalLLaMA, 2023b;a). For fine-tuning with Linear PI, we set the global batch size to 64 and tuning length to 8K, which follows Chen et al. (2023). For fine-tuning with 16K context length, we set the global batch size 32. When we fine-tune LLaMA2 7B with the last 36 dimensions being cut off, discussed in Section 3.4, we set the softmax scale $\sqrt{92}$, the square root of updated dimension size of q_t, k_s , and keep other setups the same. Except for the position embedding, we do not modify the architecture of LLaMA2 (Touvron et al., 2023b).

³<https://github.com/OpenLMLab/collie>

⁴<https://opencompass.org.cn/>

	NQ	TriviaQA	MMLU	SuperGLUE
LLaMA2 7B (Touvron et al., 2023b)	19.06	54.09	46.78	52.99
NTK fixed $\alpha = 8$ (LocalLLaMA, 2023b)	19.89	53.16	39.36	51.84
Fine-tuned	20.14	55.14	47.73	55.68
Linear PI $\lambda = 2$ (Chen et al., 2023)	18.59	53.32	47.27	54.71
Scaling RoPE base=2608	16.01	51.21	37.07	53.75
Scaling RoPE base=1304	17.70	51.27	33.46	54.55
Scaling RoPE base=652	17.12	48.90	33.43	54.47
Scaling RoPE base=500	<u>20.36</u>	52.43	38.70	53.42
Scaling RoPE base=40000	17.34	<u>55.27</u>	<u>47.34</u>	54.85
Scaling RoPE base=80000	17.81	54.72	<u>46.87</u>	54.88
Scaling RoPE base=160000	18.98	55.28	46.24	55.02
Scaling RoPE base=400000	20.39	54.91	46.01	<u>55.55</u>
Scaling RoPE base=600000	19.20	54.75	46.39	54.77
Scaling RoPE base=1000000	19.50	54.76	46.29	55.09
LLaMA2 13B (Touvron et al., 2023b)	24.82	60.69	<u>55.77</u>	54.22
NTK fixed $\alpha = 8$ (LocalLLaMA, 2023b)	22.77	60.81	52.04	53.99
Fine-tuned	26.68	61.57	55.78	55.86
Linear PI $\lambda = 2$ (Chen et al., 2023)	25.96	60.78	54.37	53.58
Scaling RoPE base=2608	20.00	59.62	55.06	55.93
Scaling RoPE base=1304	13.60	59.12	53.59	56.14
Scaling RoPE base=652	15.98	58.67	50.49	<u>56.54</u>
Scaling RoPE base=500	11.50	57.54	47.98	56.71
Scaling RoPE base=40000	26.87	60.80	55.26	55.77
Scaling RoPE base=80000	26.76	61.25	55.32	55.55
Scaling RoPE base=160000	<u>27.06</u>	61.22	54.35	55.41
Scaling RoPE base=400000	27.53	60.94	54.01	55.04
Scaling RoPE base=600000	26.81	<u>61.28</u>	53.75	55.69
Scaling RoPE base=1000000	26.51	59.22	52.94	55.33

Table 3: Short context validation results of scaling RoPE with different base and other extrapolating methods in LLaMA2 (Touvron et al., 2023b). The first box is the result of 7B size and the second box is those of 13B version. LLaMA2 7B or 13B corresponds to the original result as well as that of Dynamic NTK (LocalLLaMA, 2023a) since Dynamic NTK does not affect the model within the training context. For fixed NTK (LocalLLaMA, 2023b), $\alpha = 8$, while for linear position interpolation, $\lambda = 2$, the ratio between the fine-tuning context length and the pre-training context length. The final score of MMLU and SuperGLUE is the average score of all corresponding subtasks.

B.2 SHORT CONTEXT VALIDATION

Next, we validate whether Scaling RoPE, namely the fine-tuning RoPE with different bases on original context length, has a side effect on LLM. We use short context tasks, such as NQ (Kwiatkowski et al., 2019), TriviaQA (Joshi et al., 2017), MMLU (Hendrycks et al., 2020) and SuperGLUE (Wang et al., 2019), to assess whether the model loses the knowledge acquired in pre-training after the fine-tuning process. The results obtained are shown in the Table 3. We find that the fine-tuning results, though slightly lower compared to the original LLaMA2 7B and 13B (Touvron et al., 2023b) and directly fine-tuned version, are still acceptable. In fact, in some tasks, the performance of Scaling-RoPE even appears to be better than before. This indicates that our approach not only enhances the extrapolation performance but also does not significantly harm the inherent knowledge of LLM.

C EXTENSION

Concerning that the current extrapolation methods in the fine-tuning phase dominantly apply longer tuning contexts, we propose an extended version of the scaling law of RoPE-based extrapolation.

Theorem 3. (Extended Scaling Law of RoPE-based Extrapolation) For RoPE-based LLMs pre-trained with context length T_{train} and critical dimension d_{extra} , if we adjust the base to β and then

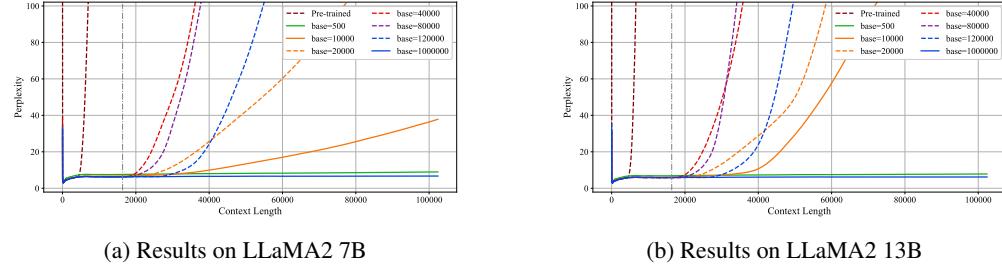


Figure 9: Perplexity of RoPE fine-tuned with 16K context length and smaller or larger bases on the validation data of Books3 (Presser, 2020). Surprisingly, as the base increases from 500 to 1000000, the extrapolation capability first becomes weaker and then becomes stronger.

conduct fine-tuning with context length $T_{\text{tune}} \geq T_{\text{train}}$, the extrapolation performance of RoPE-based LLMs will get improved. Importantly, there exists a **critical base** β_0 decided by T_{tune} and T_{train} .

$$\beta_0 = 10000^{\log \frac{T_{\text{train}}}{2\pi} \frac{T_{\text{tune}}}{2\pi}}. \quad (13)$$

If $\beta > \beta_0$, the extrapolation upper bound is decided by β and d_{extra} as follows:

$$T_{\text{extra}} = 2\pi \cdot \beta^{d_{\text{extra}} \cdot \frac{1}{d}} = 2\pi \cdot \beta^{\left\lceil \frac{d}{2} \log_{10000} \frac{T_{\text{train}}}{2\pi} \right\rceil \cdot \frac{2}{d}}. \quad (14)$$

Otherwise, the extrapolation upper bound is T_{tune} and the critical dimension is updated satisfying

$$d'_{\text{extra}} = 2 \left\lceil \frac{d}{2} \log_{\beta} \frac{T_{\text{tune}}}{2\pi} \right\rceil \geq 2 \left\lceil \frac{d}{2} \log_{10000} \frac{T_{\text{train}}}{2\pi} \right\rceil = d_{\text{extra}}. \quad (15)$$

However, the extrapolation beyond T_{tune} will acquire further enhancement if β gets smaller. Particularly, when β is smaller than $\beta_1, \beta_2, \beta_3$ as follows, the enhancement will be more significant.

$$\beta_1 = \frac{2T_{\text{tune}}}{\pi}, \quad \beta_2 = \frac{T_{\text{tune}}}{\pi}, \quad \beta_3 = \frac{T_{\text{tune}}}{2\pi}. \quad (16)$$

Theorem 3. serves as both a combination and generalization of Theorem 1., Lemma 1. and Theorem 2. Here, the critical base is the worst base for extrapolation as well as the smallest base forcing RoPE to extrapolate depending on the feature dimensions within the critical dimension. Specifically, when $T_{\text{tune}} = T_{\text{train}}$, the critical base, $\beta_0 = 10000$, relates to the intriguing observation outlined in Section 2. Equation 14 corresponds to tuning RoPE with larger bases discussed in Section 3.3, Equation 15 corresponds to the definition of critical dimension in Section 3.2 and Equation 16 corresponds to tuning RoPE with smaller bases discussed in Section 3.1. If $T_{\text{tune}} > T_{\text{train}}$, RoPE-based LLMs can accommodate a broader context window. As illustrated in Figure 9, when fine-tuning LLaMA2 7B and 13B (Touvron et al., 2023b) within a 16K context, the max context length exceeds or equals 16K, surpassing the original LLaMA2 7B and 13B respectively, regardless of the base value.

From a periodic perspective, since both base and training length have changed, it prompts us to consider whether additional positional information has been integrated during the fine-tuning phase for feature dimensions beyond the critical dimension, namely the 92nd dimension for LLaMA2 (Touvron et al., 2023b), as indicated in Equation 5. According to the definition of the critical dimension, based on the current base β , we can calculate how many dimensions the RoPE-based LLM has where sin and cos complete a period within the tuning length. If base $\beta > \beta_0$, then \tilde{d}_{extra} , the number of dimensions that cover a period during fine-tuning have already been able to traverse a complete period during pre-training, given that

$$\begin{aligned} \tilde{d}_{\text{extra}} &= 2 \left\lceil \frac{d}{2} \log_{\beta} \frac{T_{\text{tune}}}{2\pi} \right\rceil \leq 2 \left\lceil \frac{d}{2} \log_{\beta_0} \frac{T_{\text{tune}}}{2\pi} \right\rceil = 2 \left\lceil \frac{d}{2} \log_{10000} \log_{\frac{T_{\text{train}}}{2\pi}} \frac{T_{\text{tune}}}{2\pi} \right\rceil \\ &= 2 \left\lceil \frac{d}{2} \frac{1}{\log_{\frac{T_{\text{train}}}{2\pi}} \frac{T_{\text{tune}}}{2\pi}} \log_{10000} \frac{T_{\text{tune}}}{2\pi} \right\rceil = 2 \left\lceil \frac{d}{2} \frac{\log_{\frac{T_{\text{tune}}}{2\pi}} \frac{T_{\text{train}}}{2\pi}}{\log_{\frac{T_{\text{tune}}}{2\pi}} 10000} \right\rceil. \\ &= 2 \left\lceil \frac{d}{2} \log_{10000} \frac{T_{\text{train}}}{2\pi} \right\rceil = d_{\text{extra}} \end{aligned}$$

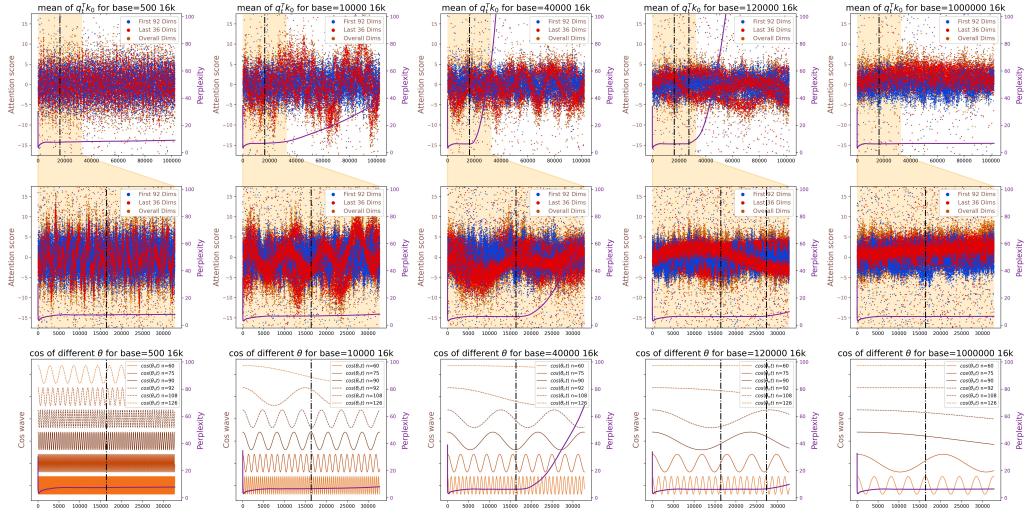


Figure 10: The relation between attention scores in first 92 and last 36 dimensions with the extrapolation performance in LLaMA 7B (Touvron et al., 2023b) evaluated or fine-tuned with different bases at 16K context length. The meaning of each row is the same as that in Figure 6 except that the second row highlights the changes in the first 32K tokens instead of the first 16K tokens.

Therefore, the critical dimension remains unchanged. Referring to Theorem 2 in Section 3.3, we can calculate the extrapolation upper bound based on the updated base and the original critical dimension as Equation 14, exactly the same as Equation 5. For LLaMA2 (Touvron et al., 2023b) fine-tuned with a 16K context, as illustrated in Figure 9, the critical base is 71738, given Equation 13. For bases greater than 71738, such as 80000, 120000, and 1000000, their extrapolation upper bounds surpass 16K and the larger base corresponds to a longer context, corroborating our theoretical framework.

If base $\beta \leq \beta_0$, then during the fine-tuning phase, the number of dimensions able to complete a period surpasses the original critical dimension, so the critical dimension is updated as Equation 15. Besides, since this dimension depends on the fine-tuning length, the extrapolation upper bound is still constrained within the fine-tuning length T_{tune} . However, if β is so small that the input of every $\cos(t-s)\theta_n, \sin(t-s)\theta_n$ can span values from 0 to $\pi/2, \pi$, or 2π within the fine-tuning length T_{tune} , as indicated by Equation 16, similar to Theorem 1 in Section 3.1, then the extrapolation performance will get further improved, marked by a more stable perplexity growth curve. For LLaMA2 (Touvron et al., 2023b) fine-tuned with a 16K context, as shown in Figure 9, for bases smaller than the critical base 71738, such as 60000, 20000, 10000, and 500, the performance curves become progressively more stable. Among them, although $\beta = 10000$ performs poorly in fine-tuning at the original context length, the performance gets significantly improved this time because the inputs of cos or sin have traversed to $\pi/2$ within the 16K context length. When $\beta = 500$, LLaMA2 achieved a similar terrific performance as $\beta = 1000000$, namely the design of Code LLaMA (Rozière et al., 2023), a context length with at least 100K tokens, breaking the curse of entropy explosion mentioned in Han et al. (2023). Since there exists an upper bound for extrapolation based on $\beta = 1000000$, RoPE tuned with base 500 on 16K context length has the potential to extrapolate to an infinite context, thus answering Q3 in the Introduction.

Similarly, we also use scatter plots to visualize the fluctuation of attention scores for different bases after fine-tuning with a 16K context as illustrated in Figure 10. For base 500, given its exposure to enough fluctuations during the training phase, the perplexity curve remains notably stable. For base 10000, it is clear that the fluctuation of attention scores in the last 36 dimensions is somewhat limited, leading to a noticeable improvement in extrapolation performance given Equation 16. For base 40000, the position information acquired in the fine-tuning phase shrinks further as the base increases. For base 120000, the critical dimension goes back to 92 dimensions, and the extrapolation performance is governed by the first 92 dimensions. Remarkably, the extrapolation upper bound given Equation 14 matches the maximum supported context length. For base 1000000, the period of the first 92 dimensions is further extended, corresponding to a context length expanding beyond

100K. Eventually, based on the above interpretation, we validate the correctness of Theorem 3. and provide a unified framework to explain the extrapolation of RoPE-based LLM with arbitrary base and fine-tuning context length.

D DISCUSSION

Besides, we discuss the instructive value of our theory for other extrapolation strategies focused on achieving longer context during the testing phase. These methods are still necessary given two facts. On one hand, the performance of RoPE with a smaller base is still left behind compared with RoPE with much larger bases, such as 1000000, as shown in Figure 1. On the other hand, for RoPE with a base that is not large enough, it still can not extrapolate to a context of 100K or longer as shown in Figure 2. In order to further enhance RoPE’s adaptability to a longer context, whatever the base value is, we discuss the effect of two complementary methods in the inference phase, log-scaled attention (Su, 2023b) and dynamic-scaled RoPE (LocalLLaMA, 2023a) on RoPE with different bases.

Log-scaled Attention as shown in Equation 17, is a classic technique originally raised in Chiang & Cholak (2022) and currently applied in RoPE-based extrapolation (Su, 2023b; Alibaba, 2023). It involves multiplying the original attention matrix by the logarithm of the current inference length t . Traditionally, the base of the logarithm is training length T_{train} . However, given Equation 14 in Theorem 3., the attention score within the max supported context length T_{extra} is reliable. So we take T_{extra} as the logarithm base and set the lower limit for the logarithmic correction value as 1, meaning that no additional log scale is required within the extrapolation upper bound.

$$\mathbf{A}_{t,s} = \operatorname{Re} \left[\sum_{n=0}^{d/2-1} \tilde{q}_t^{(n)} \tilde{k}_s^{(n)*} e^{i(t-s)\theta_n} \right] \cdot p_t. \quad (17)$$

$$p_t = \max(1, \log_{T_{\text{extra}}} t)$$

Besides Log-scaled attention, window method, such as sliding window and its variant, is also a widely accepted strategy for extrapolation, used in inference or evaluation (Press et al., 2021; Sun et al., 2022). Compared with the above strict window-based method, we follow the xPos method proposed in Sun et al. (2022), shown in Equation 18, originally used in the pre-training phase. In this work, we regard this method as a soft sliding window used in the inference phase as use it as a further complement to the log-scaled method. Still, we do little modification besides using the T_{extra} as the denominator instead of the original denominator T_{train} .

$$\mathbf{A}_{t,s} = \operatorname{Re} \left[\sum_{n=0}^{d/2-1} \tilde{q}_t^{(n)} \tilde{k}_s^{(n)*} \zeta_n^{\frac{t-s}{T_{\text{extra}}}} e^{i(t-s)\theta_n} \right]. \quad (18)$$

$$\zeta_n = \frac{\gamma + 2n/d}{\gamma + 1}, \quad n = 0 \dots \frac{d}{2} - 1, \quad \gamma = 0.4$$

Dynamic-scaled RoPE namely Dynamic NTK (LocalLLaMA, 2023a) is a widely used extrapolation. Here, we only do two little modifications. One is to change the base 10000 in vanilla RoPE (Su et al., 2021) with the base scaled in the fine-tuning phase, β . The other is still to replace the T_{train} in Equation 7 with T_{extra} we derive given Equation 14 in Theorem 3.

$$\text{for } \mathbf{A}_{t,s}, \theta_n = (\beta \cdot \alpha_t)^{-2n/d}, \text{ where } \alpha_t = \max \left(1, 2^{\lceil \log_2 \frac{t}{T_{\text{extra}}} \rceil + 1} - 1 \right). \quad (19)$$

We experiment with these two methods on LLaMA2 7B (Touvron et al., 2023b) and get the results as shown in Figure 11. Figure 11a shows the results on LLaMA2 based on RoPE with base 10000. It is clear that both log-scaled attention almost does nothing for pre-trained and fine-tuned LLaMA2, but show great improvement for fine-tuning LLaMA2 with cutting the last 36 dimensions in $\mathbf{q}_t, \mathbf{k}_s$ off. This phenomenon further proves that the fluctuation coming from the dimensions of $\mathbf{q}_t, \mathbf{k}_s$ beyond the critical dimension is the root cause of the extrapolation problem of RoPE-based LLM.

Figure 11b shows the results on LLaMA2 based on RoPE with bases smaller than 10000. There is a clear trend that with the reduction of the base value, the improvement obtained from the log-scaled

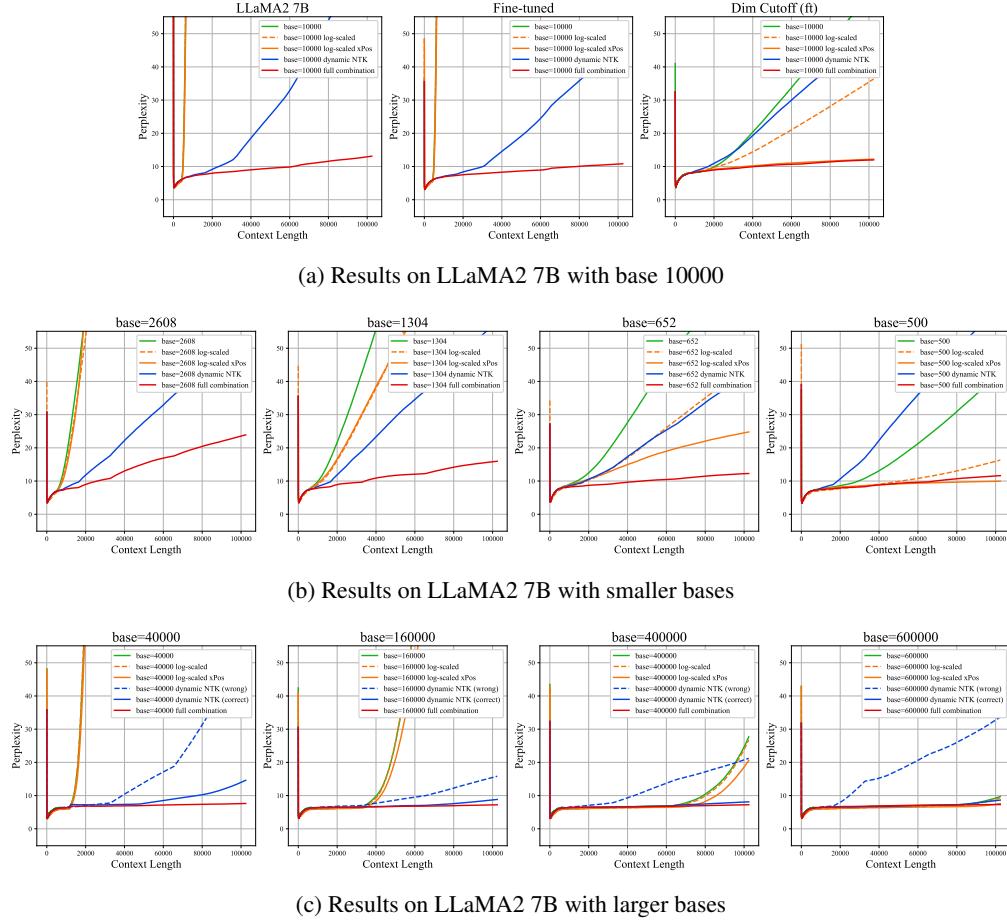


Figure 11: Perplexity on the validation data from Books3 dataset (Presser, 2020) of LLaMA 7B (Touvron et al., 2023b) based on RoPE with different bases enhanced with log-scaled attention, xPos and dynamic-scaled RoPE and their combination. Here, *wrong* means calculating α_t in Dynamic NTK (LocalLLaMA, 2023a) or p_t in (Su, 2023b) with training length T_{train} , while *correct* means using extrapolation upper bound T_{extra} instead of T_{train} .

attention is more dominant while the effect of Dynamic NTK shrinks gradually. For RoPE with base 500, the perplexity curve of log-scaled RoPE is flat enough, indicating the extrapolation capability to support 100K context length. On the contrary, Dynamic NTK shows a clear side effect. Hence, the position information learned in the training phase is reliable enough for LLM to extrapolate further, corresponding to Theorem 1. in Section 3.1.

Figure 11c shows the results on LLaMA2 based on RoPE with bases larger than 10000, such as 40000, 160000, 400000, and 600000. We do not test the performance of two methods on RoPE with base 1000000, since it already achieved the context length of 100K. Here, we enable RoPE with bases larger than 10000 and smaller than 1000000 to extrapolate beyond the context length of 100K in the inference phase. For RoPE with larger bases, the improvement of extrapolation performance obtained from Dynamic NTK is more remarkable. The working principle of Dynamic NTK has been visualized in Figure 8 and discussed in Section 3.4.

Besides, replacing the T_{train} with d_{extra} becomes significantly important for RoPE with larger bases. For example, if Dynamic NTK is carried out based on T_{train} , the improvement will be limited and even destroyed when the base is large enough like 400000 and 600000. This phenomenon proves the guidance value of this work for other extrapolation works. In conclusion, for a base smaller than β_3 defined in Equation 1, every dimension has learned complete positional information. Then the log-scaled method is sufficient to enhance extrapolation. For a base larger than β_0 , namely 10000 for

tuning on the original context, Dynamic NTK in the correct way is a good helper for extrapolation to a much longer context.

E TEXT CONTINUATION

Finally, we execute a text continuation experiment. For LLaMA2 7B(Touvron et al., 2023b) fine-tuned with various base values and fine-tuning lengths, we provide a context of 32K tokens from Books3(Presser, 2020) and prompt the model to continue writing. The resulting texts, after cleaning the special characters, are shown in Figure 12. Remarkably, whether the base is set at 500 or 1000000, the generated text remains coherent, grammatically accurate, and logically consistent. For base 500, extended tuning length or incorporating log-scaled attention Su (2023b) in the testing phase yields improved continuation results. Given the strict extrapolation upper bound for base 1000000, there is a compelling case that a model fine-tuned with base 500 possesses an infinite extrapolation potential.

LIMITATION

In this work, our primary objective is to elucidate the mechanisms by which RoPE, using either smaller or larger bases, enhances extrapolation performance. Furthermore, there remains a need to rigorously assess our method on extended context data. Additionally, it needs additional efforts to integrate whether there exist upper and lower bounds for adjusting the base. Finally, since the base value and pre-training context length is the key to extrapolation, there might be an optimal pair of base and training length for RoPE-based LLM pre-training to be discovered.

Context: THIS BOOK IS FOR INFORMATIONAL PURPOSES ONLY. IT IS NOT INTENDED TO BE A SUBSTITUTE FOR CONSULTATION WITH A PROFESSIONAL HEALTHCARE PRACTITIONER. They invited airplane pilots with an average age of fifty-two who were interested in participating and evaluated performance in flight simulations. Every three minutes, the pilots were peppered with complex commands and codes that had to be remembered and implemented while under stressful flying conditions. After seven simulations, the study subjects were given Aricept for

Reference: a month and then retested under equally demanding but new simulations. The results showed that they performed better in the later round of evaluations, and the authors stated that Aricept "appears to have beneficial effects on retention of training on complex aviation tasks in nondemented older adults," and these findings were ultimately published in Neurology. There is already a major pharmacologic market for physical performance-enhancing drugs. With an aging population, interest in medications for cognitive enhancement will likely grow.

Output (LLaMA2 7B): OOOOOOOOOOOOOOOOOOOOOOOOOOOOOOO

Output (base=500): the purpose of improving their cognitive performance. Neuro-Geek: Alzheimer's Drugs for the Healthy? While most adults will not develop Alzheimer's, many adults will have mild cognitive impairments that affect how they perform in daily life and work. In the sector of the scientific community focused on developing drugs to improve the symptoms associated with dementia, there is an understanding that those same drugs may be used by healthy people with the goal of maintaining cognitive acuity or even, in certain circumstances, improving mental agility above their established baseline.

Output (base=1000000): a month. The results were clear: the drug improved their performance. The study was published in 2015, and the authors concluded that "the results suggest that the drug may be able to improve cognitive performance in healthy individuals." The study was small, and the results are not yet conclusive. But it is a start. Prescription Antidepressants The first antidepressants were developed in the 1950s, and they have been used to treat depression ever since. But in the 1990s, scientists began to wonder if they might also be useful for improving cognitive functioning.

Output (base=500 16K): a week. The drug is a cholinesterase inhibitor, which means it blocks the breakdown of acetylcholine, a neurotransmitter that helps regulate memory and learning. The results were striking. The pilots who took the drug performed better than those who didn't. They were able to remember more of the complex commands and codes, and they were able to implement them more quickly. The study authors concluded that the drug "may be useful in improving cognitive performance in healthy individuals." But there's a catch. The drug is not approved for use in healthy people. It's only approved for people with Alzheimer's disease.

Output (base=1000000 16K): a month, then switched to a placebo for another month. The results were striking. The pilots who took the drug improved their performance by 20 percent, while those who took the placebo got worse. The study was small, and the results were not replicated in a larger study. But the idea that healthy people might be able to improve their cognitive performance with drugs designed to treat dementia is intriguing. Prescription Antidepressants The most widely prescribed class of antidepressants, selective serotonin reuptake inhibitors (SSRIs), are also used to treat anxiety, obsessive-compulsive disorder, and panic disorder.

Output (base=500 16K log-scaled): a week. The drug is a cholinesterase inhibitor, which means it blocks the breakdown of acetylcholine, a neurotransmitter that is critical for memory and learning. The results were striking. The pilots who took the drug performed better than those who didn't. They were able to remember more of the complex commands and codes and to implement them more quickly. But the study also had a major flaw. The pilots were not randomly assigned to take the drug or not. They were all given the drug because they were already healthy. The study's authors concluded that the drug improved performance in pilots who were already healthy. But that's not what the study actually showed.

Figure 12: Text continuation after a 32K-length context from Books3 (Presser, 2020) of LLaMA2 7B(Touvron et al., 2023b) tuned with base 500 or 1000000 and context length 4K or 16K. It shows that our methods could still produce cohesive and fluid output text in a much longer context.