Base of RoPE Bounds Context Length

Mingyu Xu[1]*, **Xin Men**[1]*, **Bingning Wang**[1]†, **Qingyu Zhang**[2], **Hongyu Lin**[2], **Yaojie Lu**[2], **Xianpei Han**[2] **and Weipeng Chen** [1]
[1] Baichuan Inc.
[2] Chinese Information Processing Laboratory
Institute of Software, Chinese Academy of Sciences
{menxin,xumingyu,daniel}@baichuan-inc.com
{zhangqingyu2024,hongyu,yaojie,xianpei}@iscas.ac.cn

Abstract

Position embedding is a core component of current Large Language Models (LLMs). Rotary position embedding (RoPE), a technique that encodes the position information with a rotation matrix, has been the de facto choice for position embedding in many LLMs, such as the Llama series. RoPE has been further utilized to extend long context capability, which is roughly based on adjusting the *base* parameter of RoPE to mitigate out-of-distribution (OOD) problems in position embedding. However, in this paper, we find that LLMs may obtain a superficial long-context ability based on the OOD theory. We revisit the role of RoPE in LLMs and propose a novel property of long-term decay, deriving that the *base of RoPE bounds context length*: there is an absolute lower bound for the base value to obtain certain context length capability. Our work reveals the relationship between context length and RoPE base both theoretically and empirically, which may shed light on future long context training.

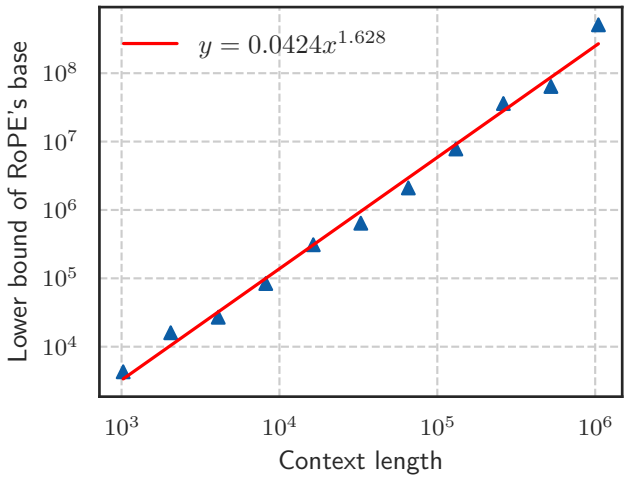


Figure 1: Context length and its corresponding lower bound of RoPE's base value.

*Equal contribution. Order determined by swapping the one in [1]
†Corresponding author.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

https://doi.org/10.52202/079017-2773

1 Introduction

In the past few years, large language models have demonstrated surprising capabilities and undergone rapid development. By now, LLMs have been widely applied across various domains, including chatbots, intelligent agents, and code assistants [2, 3]. The Transformer [4], based on the attention mechanism, has been the most popular backbone of LLMs due to its good performance and scaling properties [5]. One of the key component modules in the Transformer is position embedding, which is introduced to embed positional information that is vital for processing sequential data. Rotary position embedding (RoPE), which encodes relative distance information in the form of absolute position embedding [6], has been a popular choice and applied in many LLMs [7, 8, 9].

RoPE introduces no training parameters and shows improvement in language modeling and many other tasks [6, 10]. One reason that RoPE is widely used is its ability for context length extrapolation [11, 12], which extends the context length of a trained LLM without expensive retraining. In practice, many works [7, 13, 14] have successfully extended the window length by simply increasing base value, the only one hyper-parameter in RoPE, and fine-tuning on long texts.

The reasons behind the success of these long context extensions are often explained as avoiding out-of-distribution (OOD) rotation angles [15, 16] in RoPE, meaning the extended context length (OOD) can be mapped to the in-distribution context length that has been properly trained. Based on the OOD theory, a recent study [15] finds that a smaller base can mitigate OOD and is beneficial for the model's ability to process long contexts, which inspires us to further study the relationship between the base of RoPE and the length of context the model can process.

In this paper, we find that the model may show superficial long context capability with an inappropriate RoPE base value, in which case the model can only preserve low perplexity but loses the ability to retrieve long context information. We also show that the out-of-distribution (OOD) theory in position embedding, which motivates most length extrapolation works [11, 12, 15], is insufficient to fully reflect the model's ability to process long contexts. Therefore, we revisit the role of RoPE in LLMs and derive a novel property of long-term decay in RoPE: the ability to pay more attention to similar tokens than random tokens decays as the relative distance increases. While previous long context works often focus on the relative scale of the RoPE base, based on our theory, we derive an absolute lower bound for the base value of RoPE to obtain a certain context length ability, as shown in Figure 1. To verify our theory, we conducted thorough experiments on various LLMs such as Llama2-7B [17], Baichuan2-7B [8] and a 2-billion model we trained from scratch, demonstrating that this lower bound holds not only in the fine-tuning stage but also in the pre-training stage.

We summarize the contributions of the paper as follows:

- **Theoretical perspective**: we derive a novel property of long-term decay in RoPE, indicating the model's ability to attend more to similar tokens than random tokens, which is a new perspective to study the long context capability of the LLMs.

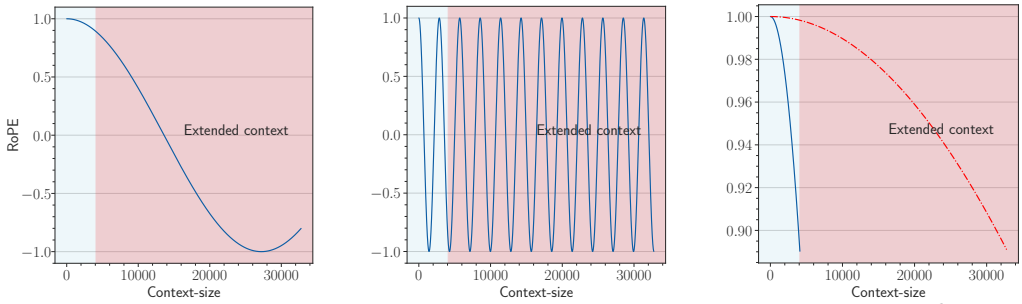
- **Lower Bound of RoPE's Base**: to achieve the expected context length capability, we derive an absolute lower bound for RoPE's base according to our theory. In short, the base of RoPE bounds context length.

- **Superficial Capability**: we reveal that if the RoPE's base is smaller than a lower bound, the model may obtain superficial long context capability, which can preserve low perplexity but lose the ability to retrieve information from long context.

2 Background

In this section, we first introduce the Transformer and RoPE, which are most commonly used in current LLMs. Then we discuss long context methods based on the OOD of rotation angle theory.

https://doi.org/10.52202/079017-2773



(a) base=1e4 (b) base=500 (c) base=$b \cdot s^{\frac{d}{d-2}}$

Figure 2: An illustration of OOD in RoPE when we extend context length from 4k to 32k, and two solutions to avoid the OOD. We show the last dimension as it is the lowest frequency part of RoPE, which suffers OOD mostly in extrapolation. (a) For a 4k context-length model with base value as 1e4, when we extend the context length to 32k without changing the base value, the context length from 4k to 32k is OOD for RoPE (red area in the figure). (b) OOD can be avoided with a small base value like 500 [15], since the full period has been fitted during fine-tuning stage. (c) We set base as $b \cdot s^{\frac{d}{d-2}}$ from NTK [11].The blue line denotes the pre-training stage (base=1e4) and the red dashed line denotes the fine-tuning stage (base=$b \cdot s^{\frac{d}{d-2}}$), we can observe that the RoPE's rotation angle of extended positions is in-distribution.

2.1 Attention and RoPE

The LLMs in current are primarily based on the Transformer [4]. The core component of it is the calculation of the attention mechanism. The naive attention can be written as:

$$A_{ij} = q_i^T k_j \tag{1}$$

$$\text{ATTN}(X) = \text{softmax}(A/\sqrt{d})\, v, \tag{2}$$

where $A \in R^{L \times L}$ $q, k, v \in R^d$. Position embedding is introduced to use the order of the sequence in attention.

RoPE [6] implements relative position embedding through absolute position embedding, which applies rotation matrix into the calculation of the attention score in Eq. 1, which can be written as:

$$A_{ij} = (R_{i,\theta} q_i)^T (R_{j,\theta} k_i) = q_i^T R_{j-i,\theta} k_j = q_i^T R_{m,\theta} k_j, \tag{3}$$

where $m = j - i$ is the relative distance of i and j, $R_{m,\theta}$ is a rotation matrix denoted as:

$$\begin{bmatrix} cos(m\theta_0) & -sin(m\theta_0) & 0 & 0 & \cdots & 0 & 0 \\ sin(m\theta_0) & cos(m\theta_0) & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & cos(m\theta_1) & -sin(m\theta_1) & \cdots & 0 & 0 \\ 0 & 0 & sin(m\theta_1) & cos(m\theta_1) & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & cos(m\theta_{d/2-1}) & -sin(m\theta_{d/2-1}) \\ 0 & 0 & 0 & 0 & \cdots & sin(m\theta_{d/2-1}) & cos(m\theta_{d/2-1}) \end{bmatrix} \tag{4}$$

Generally, the selection of rotation angles satisfies $\theta_i = base^{-2i/d}$, the typical base value for current LLMs is 10,000.

2.2 OOD theory of relative rotation angle

Based on RoPE, researchers have proposed various methods to extend the long context ability of LLMs, among which representatives are PI [12] and NTK-series (NTK-aware [18], YaRN [11], and Dynamical-NTK [19]). Those methods depend on the relative scale $s = T_{\text{new}}/T_{\text{origin}}$, where T_{origin} is the training length of the original pre-trained model and T_{new} is the training length in long-context fine-tuning.

PI PI directly interpolates the position embedding, and the calculation of A_{ij} becomes:

$$A_{ij} = (R_{i/s} q_i)^T (R_{j/s} k_i) = q_i^T R_{(j-i)/s} k_j = q_i^T R_{m/s} k_j, \tag{5}$$

In other words, the position embedding of the token at position i in pre-training becomes i/s in fine-tuning, ensuring the position embedding range of the longer context remains the same as before.

https://doi.org/10.52202/079017-2773

NTK-series The idea is that neural networks are difficult to learn high-frequency features, and direct interpolation can affect the high-frequency parts. Therefore, the NTK-aware method achieves high-frequency extrapolation and low-frequency interpolation by modifying the base value of RoPE. Specifically, it modifies the base b of the RoPE to:

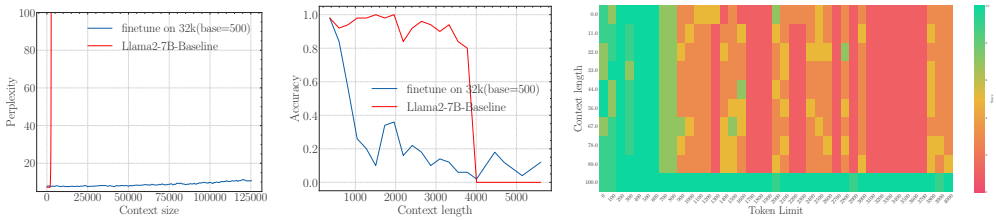
$$b_{\text{new}} = b\, s^{\frac{d}{d-2}}. \tag{6}$$

The derivation of this expression is derived from $T_{\text{new}} b_{\text{new}}^{-\frac{d-2}{d}} = T_{\text{origin}} b^{-\frac{d-2}{d}}$ to ensure that the lowest frequency part being interpolated.

A recent study [15] proposes to set a much smaller base (e.g. 500), in which case $\theta_i = base^{-\frac{2i}{d}}$ is small enough and typical training length (say 4,096) fully covers the period of $\cos(t-s)\theta_i$, so the model can obtain longer context capabilities.

One perspective to explain current extrapolation methods is the OOD of rotation angle [15, 16]. If all possible values of $\cos(t-s)\theta_i$ have been fitted during the pre-training stage, OOD would be avoided when processing longer context. Figure 2 demonstrates how these methods avoid OOD of RoPE.

3 Motivation



(a) Perplexity (b) Long-eval [20] (c) Needle in Haystack [21]

Figure 3: The superficial long context capability of avoiding OOD by the smaller base. Following the recent work [15], we fine-tune Llama2-7B with a small base (500) to a context length of 32k.

Recent advancements in long-context language models have seen widespread adoption of NTK-based methods [7, 13, 14]. However, a curious trend has emerged: practitioners often employ significantly larger base values than those originally suggested by NTK-aware approaches. This discrepancy raises critical questions about the efficacy of current theoretical frameworks. Why do practitioners deviate from the recommendations of NTK-based methods? Is the out-of-distribution (OOD) theory underlying these methods insufficient to unlock long-context capabilities fully?

On the other hand, recent research [15], driven by OOD theory, proposes using a much smaller base for RoPE to extend context length. However, our findings, as illustrated in Figure 3, suggest that this approach may only provide superficial long-context capability[22]. While achieving low perplexity even at 128k context length (explicable by OOD theory), the model fails to retrieve relevant information for context lengths as short as 1kwell below its pre-trained length. The observation suggests that the small base determined by OOD theory can't unlock true long-context capability.

These phenomena motivate us to delve deeper into the relationship between RoPE's base and context length. To address the gap between OOD theory and our observations, we conduct a theoretical exploration in the next section, aiming to uncover the underlying mechanisms of effective long-context modeling.

4 Theory Perspective

For attention mechanism in language modeling, we have the following desiderata:

Desiderata 1 ***The closer token gets more attention****: the current token tends to pay more attention to the token that has a smaller relative distance.*

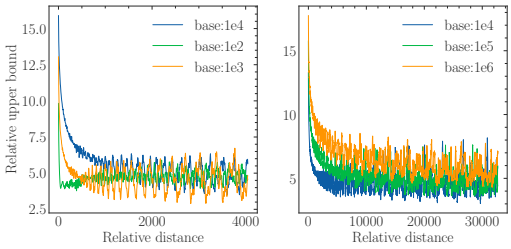


Figure 4: The upper bound of attention score with respect to the relative distance.

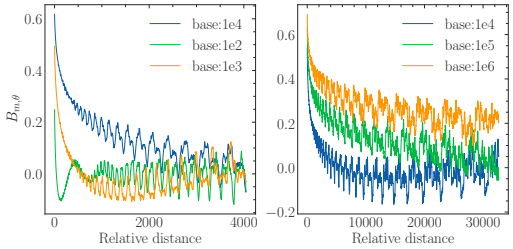


Figure 5: The ability to attend more to similar tokens than random tokens.

Desiderata 2 ***The similar token gets more attention****: the token tends to pay more attention to the token whose key value is more similar to the query value of the current token.*

Then we examine the desiderata when we apply RoPE to the attention mechanism in LLMs.

4.1 Long-term Decay of Upper Bound of Attention Score

For Desiderata 1, the property of RoPE makes the model attend more to closer tokens. This kind of long-term decay has been thoroughly discussed in previous work [6, 23]. It comes from the upper bound of attention score calculation, which can be written as:

$$\begin{aligned} |A_{ij}| = |q_i^T R_m k_j| &\leq \max_l(|h_l - h_{l+1}|) \sum_{n=1}^{d/2} |S_n| \\ &= \max_l(|h_l - h_{l+1}|) \sum_{n=1}^{d/2} |\sum_{l=0}^{n-1} e^{(j-i)\theta_l \sqrt{-1}}|, \end{aligned} \tag{7}$$

where $h_l = q_i^T[2l : l2+1]k_j[2l : 2l+1]$. Equation 7 indicates that the upper bound of the attention score $|A_{ij}|$ decays as the relative distance increases. Figure 4 shows the long-term decay curve of this upper bound, which is in accordance with previous findings [6, 23].

4.2 Long-term Decay of the Ability to Attend More to Similar Tokens than Random Tokens

In addition to the attention score's upper bound, we also find there exists another long-term decay property in RoPE: the ability to attend more to similar tokens than random tokens decays as the relative distance increases. We define the ability to attend more to similar tokens than random tokens as:

$$\mathbb{E}_{q,k^*}\left[q^T R_{m,\theta} k^*\right] - \mathbb{E}_{q,k}\left[q^T R_{m,\theta} k\right], \tag{8}$$

where $q \in R^d$ is the query vector for the current token, $k^* = q + \epsilon$ is the key value of a similar token, where ϵ is a small random variable, $k \in R^d$ is the key vector of a random token, $R_{m,\theta}$ is the rotation matrix in RoPE. The first term in Eq. 8 is the attention score of q and a similar token k^*, the second term in Eq. 8 is the attention score of q and random token k. Then we derive the following theorem:

Theorem 1 *Assuming that the components of query $q \in R^d$ and key $k \in R^d$ are independent and identically distributed, their standard deviations are denoted as $\sigma \in R$. The key $k^* = q + \epsilon$ is a token similar to the query, where ϵ is a random variable with a mean of 0. Then we have:*

$$\frac{1}{2\sigma^2}(\mathbb{E}_{q,k^*}\left[q^T R_{m,\theta} k^*\right] - \mathbb{E}_{q,k}\left[q^T R_{m,\theta} k\right]) = \sum_{i=0}^{d/2-1} \cos(m\theta_i) \tag{9}$$

The proof is shown in Appendix A. We denote $\sum_{i=0}^{d/2-1} \cos(m\theta_i)$ as $B_{m,\theta}$, and according to Theorem 1, $B_{m,\theta}$ measures the ability to give more attention to similar tokens than random tokens, which decreases as the relative distance m increases, as shown in Figure 5. For a very small base value, we can observe that the $B_{m,\theta}$ is even below zero at a certain distance, meaning the random tokens have larger attention scores than the similar tokens, which may be problematic for long context modeling.

https://doi.org/10.52202/079017-2773

Table 1: Context length and its corresponding lower bound of RoPE's base.

Context Len.	1k	2k	4k	8k	16k	32k	64k	128k	256k	512k	1M
Lower Bound	4.3e3	1.6e4	2.7e4	8.4e4	3.1e5	6.4e5	2.1e6	7.8e6	3.6e7	6.4e7	5.1e8

4.3 Base of RoPE Bounds the Context Length

To satisfy the Desiderata 2, we will get $\mathbb{E}_{q,k^*}\left[q^T R_{m,\theta} k^*\right] \geq \mathbb{E}_{q,k}\left[q^T R_{m,\theta} k\right]$. According to Theorem 1, $B_{m,\theta}$ needs to be larger than zero. Given the θ in RoPE, the context length L_θ that can be truly obtained satisfies:

$$L_\theta = \sup\{L | B_{m,\theta} \geq 0, \forall m \in [0, 1, ..., L]\} \tag{10}$$

In other word, if we follow the setting that $\theta_i = base^{-2i/d}$, in order to get the expected context length L, there is a lower bound of the base value $base_L$:

$$base_L = \inf\{base | B_{m,\theta} \geq 0, \forall m \in [0, 1, ..., L]\} \tag{11}$$

In summary, the RoPE's base determines the upper bound of context length the model can truly obtain. Although there exists the absolute lower bound, Eq. 9 and Eq. 11 are hard to get the closed-form solution since $B_{m,\theta}$ is a summation of many cosine functions. Therefore, in this paper, we get the numerical solution. Table 1 shows this lower bound for context length ranging from 1,000 to one million. In Figure 1, we plot the context length and corresponding lower bound, we can observe that as the context length increases, the required base also increases.

Note: this boundary is not very strict because the stacking of layers in LLMs allows the model to extract information beyond the single layers' range, which may increase the context length in Eq. 10 and decrease the base in Eq. 11. Notwithstanding, in Section 5 we find that the derived bound approximates the real context length in practice.

Long-term decay from different perspectives. The long-term decay in section 4.1 and section 4.2 are from different perspectives. The former refers to the long-term decay of the attention score as the relative distance increases. This ensures that current tokens tend to pay more attention to the tokens closer to them. The latter indicates that with the introduction of the rotation matrix in attention, the ability to discriminate the relevant tokens from irrelevant tokens decreases as the relative distance increases. Therefore, a large $B_{m,\theta}$, corresponding to a large base value, is important to keep the model's discrimination ability in long context modeling.

5 Experiment

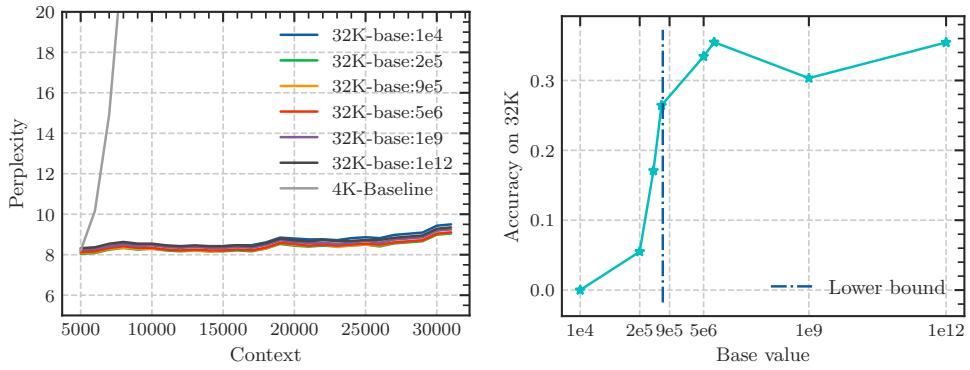
In this section, we conduct thorough experiments. The empirical result can be summarized in Table 2, the details are in the following sections.

Table 2: In Section 5, we aim to answer the following questions.

Questions	Answers
Q: Does RoPE's base bounds the context length during the fine-tuning stage?	Yes. When the base is small, it is difficult to get extrapolation for specific context length.
Q: Does RoPE's base bounds the context length during the pre-training stage?	Yes. Our proposed lower bound for RoPE's base also applies to pre-training. If we train a model from scratch with a small base but the context length is large (larger than the bounded length), the resulting model has very limited context length capabilities, meaning some of the context in pre-training is wasted.
Q: What happened when base is set smaller than the lower bound?	The model will get the superficial long context capability. The model can keep perplexity low, but can't retrieve useful information from long context.

5.1 Experiments Setup

For fine-tuning, we utilized Llama2-7B [7] and Baichuan2-7B [8], both of which are popular open-source models employing RoPE with a base of $1e4$. We utilized a fixed learning rate of 2e-5 and a



(a) Perplexity (b) Long-eval 32k

Figure 6: Fine-tuning Llama2-7B-Base on 32k context length with varying RoPE's base. Although the perplexity remains low with varying bases, the Long-eval accuracy reveals a discernible bound for the base value, below which the Long-eval accuracy declines significantly. The dotted line denotes the lower bound derived from Eq. 11 and code is provided in Appendix E

global batch size of 128 and fine-tuning for 1000 steps. For pre-training, we trained a Llama-like 2B model from scratch for a total of 1 trillion tokens. We set the learning rate to 1e-4 and adopted a cosine decay schedule, with models trained on a total of 1T tokens. The dataset we used is a subset of RedPajama [24]. More details of the experimental setup are provided in Appendix B.

Our evaluation focused on two aspects: (1) **Perplexity**: we use PG19 dataset [25] which are often used in long context evaluation; (2) **Retrieval**: in addition to perplexity, we also adopt retrieval since it represents the real long-context understanding ability of LLMs. We choose a) Long-eval benchmark from [20] and b) Needle in a haystack (NIH) [21]. The Long-eval benchmark generates numerous random similar sentences and asks the model to answer questions based on a specific sentence within the context, while the NIH requires the model to retrieve information from various positions in the long context.

5.2 Base of RoPE bounds context length in fine-tuning stages

According to Eq. 11, there is a lower bound of RoPE's base determined by expected context length. We fine-tune Llama2-7b-Base on 32k context with varying bases. As depicted in Figure 6, although the difference in perplexity between different bases is negligible, the accuracy of Long-eval varies significantly. In Figure 6b, the dotted line denotes the lower bound derived from Eq. 11, below which the Long-eval accuracy declines significantly. Additional results are provided in Appendix C. Notably, this empirically observed lower bound closely aligns with our theoretical derivation. On the other hand, we can see that $base = 2e5$ achieves the best perplexity, but the accuracy of Long-eval is very low, which indicates the limitations of perplexity in evaluating long context capabilities. We also provide the more comprehensive RULER [26]benchmark results in Appendix G.

5.3 The Base of RoPE bounds context length in pre-training stages

According to **Theorem 1** and **Eq. 11**, these constraints could also apply to the pre-training stage. To validate this, we trained a 2B model from scratch with RoPE base=100. The results, depicted in the first row of Figure 7, indicate that even though the model was trained with a context length of 4,096 tokens, it was capable of retrieving information from only the most recent approximately 500 tokens. This demonstrates that the base parameter bounds the context length during the pre-training stage as well. We define the context length from which the model can effectively retrieve information as the effective context length.

According to our theory, the effective context length can be extended as the RoPE's base increases. To validate this, we further fine-tune this 2B model on 32k context length, with RoPE's base set to 1e4, as shown in the second row of Figure 7. While the effective context length increased, it remains significantly below 32k since the effective context length bounded by base=1e4 is much smaller

https://doi.org/10.52202/079017-2773

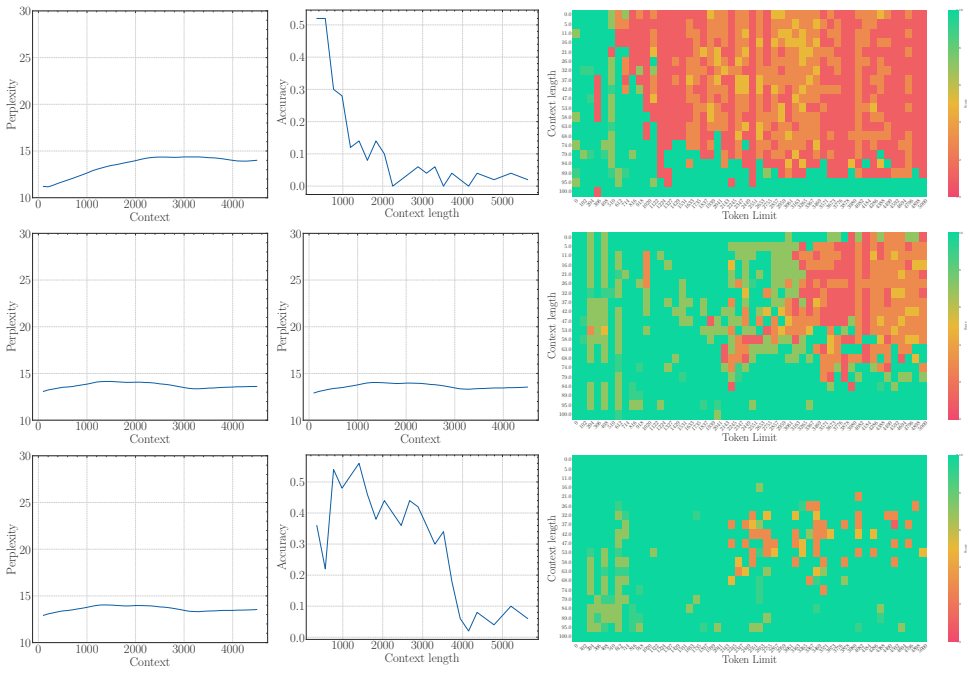


Figure 7: The first row: the results of a 2B model training from scratch with base=1e2. The second row: The results of fine-tuning the 2B model with base=1e4. The third row: The results of fine-tuning the 2B model with base=1e6.

than 32k. Further, when we increase the base to 1e6 and fine-tune the base 2B model on 32K (the third row in Figure 7), the model could obtain a larger context length than base=1e4, which is in accordance with our theory.

To further remove the influence of model size, we also fine-tuned a larger 7B model on a 32k context length with a RoPE base set to 1e4 and observed an effective context length nearly identical to that of the 2B model with the same RoPE base (see Appendix D). This is empirical proof that the effective context length is determined by RoPE's base.

5.4 Interpretation for the superficial long context capability for small base

Based on our theory and empirical observations, it is easy to explain what happens in Figure 3.

Better Extrapolation (Perplexity)? Due to the small base, $B_{m,\theta}$ can be smaller than zero as m increases, which is shown in Figure 5. The model can't attend more to similar tokens than random tokens with a large relative distance, so the model tends to focus more on nearby tokens, this will lead to a smaller empirical receptive field, even smaller than the training length. In this case, the model has a strong ability to maintain perplexity stability [27].

Worse Ability (Long-eval and NIH)! According to our previous analysis, RoPE's base bounds the context length, and the context length bounded by 500 is much lower than that bound by 10,000. Therefore, when the base is set to 500, the effective context length drops sharply, even after training on 32k context length.

5.5 OOD theory is insufficient to reveal long context capability

Section 3 mentions that methods based on the OOD theory of rotation angles may not fully reflect the long context capability. In this section, we conduct further experiments to substantiate and explain this observation. We present two methods to extend the context length of Llama2 from 4k to 32k. Both of them are devoid of OOD angles. These methods are delineated mathematically as follows:

- Method 1: $\theta_i = (5e6)^{-2i/d}$,

Table 3: The comparison of "Method 1" and "Method 2". These methods are designed carefully. They both are no OOD, but they are very different under our theory.

Method	OOD	Long-eval 15k	Long-eval 30k	numbers of m whose $B_{m,\theta} \leq 0$ 15k	numbers of m whose $B_{m,\theta} \leq 0$ 30k
Method 1	✘	0.33	0.27	0	0
Method 2	✘	0.40	0.00	97	2554

- Method 2: $\theta_i = \begin{cases} (1e4)^{-2i/128}/8, & i \geq 44 \\ (1e4 * 8^{128/88})^{-2i/128}, & i < 44. \end{cases}$

We can see from Table 3 that these two methods exhibit significantly different long context capabilities. Under the perspective of OOD rotation angle, both methods avoid OOD rotation angle, suggesting effective extrapolation. However, despite being trained on a context length of 32k, "method 2" struggles in completing the retrieval task at a context length of 32k. This phenomenon is beyond the scope which the OOD theory can explain.

Under our perspective, "method 2" is severely violating $B_{m,\theta} \geq 0$ when $m \in [15k, 30k]$, thereby impeding its ability to achieve long-context discrimination. We speculate that the model may achieve better extrapolation in the fine-tuning stage if the base is sufficiently large to surpass a lower bound and avoid OOD of rotation angles.

6 Related Work

Position embedding. Since its introduction, Transformer [4] has achieved remarkable results in the field of natural language processing. To make full use of the order of sequence, researchers have introduced position embedding. The earliest position embedding was based on sinusoidal functions [4] for absolute positions, learnable absolute position embedding [28] and many variants [29, 30] were proposed. Nevertheless, absolute position embedding has difficulties in extending directly to texts longer than the training length. Subsequently, researchers proposed relative position embedding methods [31, 32]. With the development of large language models, rotary position embedding and its variants [6, 23] has become widely used, such as Llama2 [7], Baichuan2 [8], Mistral-7B-[33]. A recent study reveals that no position embedding is also potential [34].

Long context learning. Implementing models with longer or even infinitely long contexts has always been an important goal in the field of natural language processing. Due to the squared complexity of the transformer model over time, a significant portion of the work focuses on improving the model structure [35, 35, 36, 37]. However, most of the work is still based on the transformer architecture. The other part of the work is aimed at reducing the computational complexity of attention itself, such as sparse attention [38] and group query attention [39]. In addition, there are also some optimizations in engineering efficiency, such as flash attention [40] and ring attention [41]. In the model inference stage, to save time and space, there are also some methods for accelerating long context, such as KV cache compression [42], etc. And the position embedding is important in extrapolation. In the process of fine-tuning, methods such as PI [12], NTK, and YARN [11] are used to change the original position embedding information. FoT [43] assigns the position information of the tokens outside the local context as the first token in the local context.

7 Limitation

In this work, we investigate the relationship between the base of RoPE and context length. Although we have derived that there exists a lower bound for the base of RoPE determined by context length, the existence of the upper bound for RoPE's base remains an open question that warrants further exploration. In addition, because of the lack of effective benchmarks for assessing long-context capabilities, the scope of long-context capabilities discussed in this paper may be limited.

https://doi.org/10.52202/079017-2773

8 Conclusion

Our work presents a comprehensive study on the role of RoPE in LLMs for effectively modeling long context. Our main contribution lies in uncovering a novel property of RoPE through theoretical analysis, demonstrating that as the relative distance between tokens increases, the model's ability to attend more to similar tokens decreases. According to our theory, we derive a lower bound for RoPE's base in accommodating to expected context lengths. Our experimental results validate that the base of RoPE bounds context length for not only fine-tuning but also the pre-training stage. Our theory offers a new perspective on understanding the functionality of RoPE in long-context modeling. By shedding light on the relationship between context length and position embedding, we hope our work could provide insights for enhancing the long context capability of LLMs.

References

[1] 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.

[2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[3] Nan Jiang, Kevin Liu, Thibaud Lutellier, and Lin Tan. Impact of code language models on automated program repair. In *2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE)*, pages 1430–1442. IEEE, 2023.

[4] 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.

[5] Yi Tay, Mostafa Dehghani, Samira Abnar, Hyung Won Chung, William Fedus, Jinfeng Rao, Sharan Narang, Vinh Q Tran, Dani Yogatama, and Donald Metzler. Scaling laws vs model architectures: How does inductive bias influence scaling? *arXiv preprint arXiv:2207.10551*, 2022.

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

[7] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023.

[8] Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, et al. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*, 2023.

[9] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

[10] Byeongho Heo, Song Park, Dongyoon Han, and Sangdoo Yun. Rotary position embedding for vision transformer. *arXiv preprint arXiv:2403.13298*, 2024.

[11] 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*, 2023.

[12] 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.

[13] Hao Liu, Wilson Yan, Matei Zaharia, and Pieter Abbeel. World model on million-length video and language with ringattention. *arXiv preprint arXiv:2402.08268*, 2024.

[14] 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.

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

[16] 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.

 https://doi.org/10.52202/079017-2773

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

[18] bloc97. Ntk-aware scaled rope allows llama models to have extended (8k+) context size without any fine-tuning and minimal perplexity degradation. `https://www.reddit.com/r/LocalLLaMA/comments/14lz7j5/ntkaware_scaled_rope_allows_llama_models_to_have/`, 2023.

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

[20] Dacheng Li*, Rulin Shao*, Anze Xie, Ying Sheng, Lianmin Zheng, Joseph E. Gonzalez, Ion Stoica, Xuezhe Ma, and Hao Zhang. How long can open-source llms truly promise on context length?, June 2023.

[21] Kamradt G. Needle in a haystack - pressure testing llms. `https://github.com/gkamradt/LLMTest_NeedleInAHaystack`, 2023.

[22] Yutong Hu, Quzhe Huang, Mingxu Tao, Chen Zhang, and Yansong Feng. Can perplexity reflect large language model's ability in long text understanding? In *The Second Tiny Papers Track at ICLR 2024*, 2024.

[23] 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.

[24] Together Computer. Redpajama: An open source recipe to reproduce llama training dataset, April 2023.

[25] Jack W Rae, Anna Potapenko, Siddhant M Jayakumar, Chloe Hillier, and Timothy P Lillicrap. Compressive transformers for long-range sequence modelling. In *International Conference on Learning Representations*, 2019.

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

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

[28] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[29] Shun Kiyono, Sosuke Kobayashi, Jun Suzuki, and Kentaro Inui. Shape: Shifted absolute position embedding for transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3309–3321, 2021.

[30] Hailiang Li, YC Adele, Yang Liu, Du Tang, Zhibin Lei, and Wenye Li. An augmented transformer architecture for natural language generation tasks. In *2019 International Conference on Data Mining Workshops (ICDMW)*, pages 1–7. IEEE, 2019.

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

[32] Guolin Ke, Di He, and Tie-Yan Liu. Rethinking positional encoding in language pre-training. In *International Conference on Learning Representations*, 2020.

[33] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023.

[34] Amirhossein Kazemnejad, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Payel Das, and Siva Reddy. The impact of positional encoding on length generalization in transformers. *Advances in Neural Information Processing Systems*, 36, 2024.

[35] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

[36] Bo Peng, Eric Alcaide, Quentin Anthony, Alon Albalak, Samuel Arcadinho, Stella Biderman, Huanqi Cao, Xin Cheng, Michael Chung, Leon Derczynski, et al. Rwkv: Reinventing rnns for the transformer era. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14048–14077, 2023.

[37] Zhen Qin, Songlin Yang, and Yiran Zhong. Hierarchically gated recurrent neural network for sequence modeling. *Advances in Neural Information Processing Systems*, 36, 2024.

[38] Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020.

[39] Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebron, and Sumit Sanghai. Gqa: Training generalized multi-query transformer models from multi-head checkpoints. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

[40] Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in Neural Information Processing Systems*, 35:16344–16359, 2022.

[41] Hao Liu, Matei Zaharia, and Pieter Abbeel. Ring attention with blockwise transformers for near-infinite context. In *NeurIPS 2023 Foundation Models for Decision Making Workshop*, 2023.

[42] Coleman Hooper, Sehoon Kim, Hiva Mohammadzadeh, Michael W Mahoney, Yakun Sophia Shao, Kurt Keutzer, and Amir Gholami. Kvquant: Towards 10 million context length llm inference with kv cache quantization. *arXiv preprint arXiv:2401.18079*, 2024.

[43] Szymon Tworkowski, Konrad Staniszewski, Mikołaj Pacek, Yuhuai Wu, Henryk Michalewski, and Piotr Miłoś. Focused transformer: Contrastive training for context scaling. *Advances in Neural Information Processing Systems*, 36, 2024.

[44] Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. Megatron-lm: Training multi-billion parameter language models using model parallelism, 2020.

[45] Amirkeivan Mohtashami and Martin Jaggi. Random-access infinite context length for transformers. *Advances in Neural Information Processing Systems*, 36, 2024.

 https://doi.org/10.52202/079017-2773

A The proof of Theorem 1

Assuming that the components of query $q \in R^d$ and key $k \in R^d$ are independent, their standard deviations are denoted as $\sigma \in R^d$ and the means are donated as $\mu \in R^d$. The key k^* similar to q is $q + \epsilon$, where ϵ is a random variable with a mean of 0. Then, we have:

$$
\begin{aligned}
&\mathbb{E}_{q,k^*} q^T R_m k^* - \mathbb{E}_{q,k} q^T R_m k \\
=&\mathbb{E}_q q^T R_m q + \mathbb{E}_{q,\epsilon} q^T R_m \epsilon - \mathbb{E}_{q,k} q^T R_m k \\
=&\mathbb{E}_q \sum_{i=0}^{d/2-1} (q_{2i}^2 \cos(m\theta_i) - q_{2i} q_{2i+1} sin(m\theta_i) + q_{2i+1} q_{2i} sin(m\theta_i) + q_{2i+1}^2 \cos(m\theta_i)) + \mathbb{E}_q q^T R_m \mathbb{E}_\epsilon \epsilon \\
&- \mathbb{E}_{q,k} \sum_{i=0}^{d/2-1} (q_{2i} k_{2i} \cos(m\theta_i) - q_{2i} k_{2i+1} sin(m\theta_i) + q_{2i+1} k_{2i} sin(m\theta_i) + q_{2i+1} k_{2i+1} \cos(m\theta_i)) \\
=& \sum_{i=0}^{d/2-1} \mathbb{E}(q_{2i}^2) \cos(m\theta_i) - \mu_{2i} \mu_{2i+1} sin(m\theta_i) + \mu_{2i} \mu_{2i+1} sin(m\theta_i) + \mathbb{E}(q_{2i+1}^2) \cos(m\theta_i)) + \mu R_m 0 \\
&- \sum_{i=0}^{d/2-1} (\mu_{2i}^2 \cos(m\theta_i) - \mu_{2i} \mu_{2i+1} sin(m\theta_i) + \mu_i \mu_{2i+1} sin(m\theta_i) + \mu_{2i+1}^2 \cos(m\theta_i)) \\
=& \sum_{i=0}^{d/2-1} (E(q_{2i}^2 + q_{2i+1}^2) - \mu_{2i}^2 - \mu_{2i+1}^2) \cos(m\theta_i) \\
=& \sum_{i=0}^{d/2-1} (\sigma_i^2 + \sigma_{i+1}^2) \cos(m\theta_i)
\end{aligned} \tag{12}
$$

Then we can get:

$$
\sum_{i=0}^{d/2-1} (\sigma_{2i}^2 + \sigma_{2i+1}^2) \cos(m\theta_i) = \mathbb{E}_{q,k^*} q^T R_m k^* - \mathbb{E}_{q,k} q^T R_m k \tag{13}
$$

And when all σ are equal, we can get:

$$
\sum_{i=0}^{d/2-1} \cos(m\theta_i) = \frac{1}{2\sigma^2} (\mathbb{E}_{q,k^*} q^T R_m k^* - \mathbb{E}_{q,k} q^T R_m k) \tag{14}
$$

B The detail setting of experiment

For training, we mainly conducted experiments on Llama2-7B [7] and Baichuan2-7B [8]. In addition, we also trained a 2B model from scratch, whose structure is the same as Baichuan2-7B-Base but with a smaller hidden size = 2048. Both training and testing are accelerated by FlashAttention-2 [40] and Megatron-LM [44]. The dataset of both fine-tuning and training from scratch is a subset of RedPajama [24]. The hyperparameters of training are listed in Appendix 4. All experiments are conducted on a cluster of 16 machines with 128 NVIDIA A100 80G.

Table 4: Training hyper-parameters in our experiments

Model	Training length	Training tokens	Batchsize	Base LR	LR decay	Weight decay
Llama2-7B-Base	32K	4B	128	2e5	constant	0
Baichuan2-7B-Base	32K	4B	128	2e5	constant	0
Our-2B-Base	4K	1T	1024	2e4	cosine	0.1

Question: Below is a record of lines I want you to remember. Each line begins with 'line <line index>' and contains a '<REGISTER_CONTENT>' at the end of the line as a numerical value. For each line index, memorize its corresponding <REGISTER_CONTENT>. At the end of the record, I will ask you to retrieve the corresponding <REGISTER_CONTENT> of a certain line index. Now the record start:

...

line swift-baby: REGISTER_CONTENT is <12821>
line dangerous-breast: REGISTER_CONTENT is <28051>
line bad-sculptural: REGISTER_CONTENT is <32916>
line flashy-college: REGISTER_CONTENT is <34027>
line voiceless-brochure: REGISTER_CONTENT is <8964>
line fast-peony: REGISTER_CONTENT is <5218>

...

Now the record is over. Tell me what is the <REGISTER_CONTENT> in line dangerous-breast? I need the number. **Answer**:

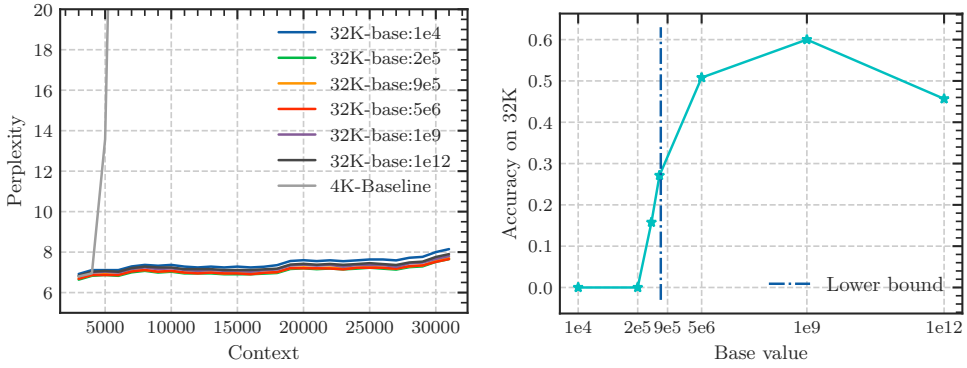
Figure 8: Long-eval sample prompt

For evaluation, we test the long context capabilities comprehensively, the benchmarks are listed below: **perplexity** on PG19 [25] test split. We evaluate the perplexity of each sample and get the mean value across samples.

Long-eval [20]. This test generates massive random similar sentences and asks the model to answer questions according to a specific sentence in the context. Because the long context consists of many similar patterns, it's more difficult to get the right answer. We find this test is harder than other long context evaluations such as Perplexity, Passkey Retrieval [45], Needle in Haystack [21]. A test sample is list in Figure 8

needle in haystack(NIH) [21]. NIH tests the long context capability not only under different context lengths but also at different positions where the correct answer is located in the context, which provides a more detailed view of the long context capability.

C Baichuan2-7B-Base: Lower bound Base of RoPE



(a) Perplexity

(b) Long-eval 32k

Figure 9: Fine-tuning Baichuan2-7B-Base on 32k context length with varying RoPE's base. Although the perplexity remains low with varying bases, the Long-eval accuracy reveals a discernible bound for the base value, below which the Long-eval accuracy declines significantly. the dotted line denotes the lower bound derived from Eq. 11.

D Long Context Test Results on Various LLMs

https://doi.org/10.52202/079017-2773

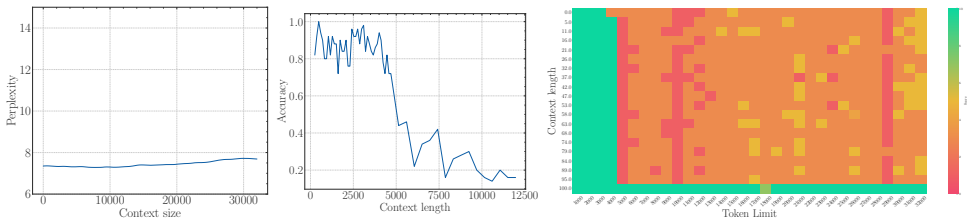


Figure 10: Llama2-7B-Base with base=1e4 fine-tuned on 32k context (original context=4096)

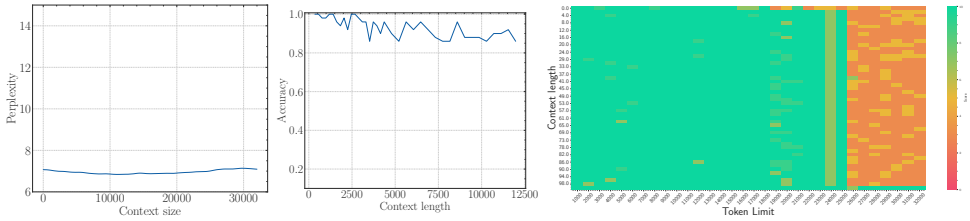


Figure 11: Llama2-7B-Base with base=2e5 fine-tuned on 32k context (original context=4096)

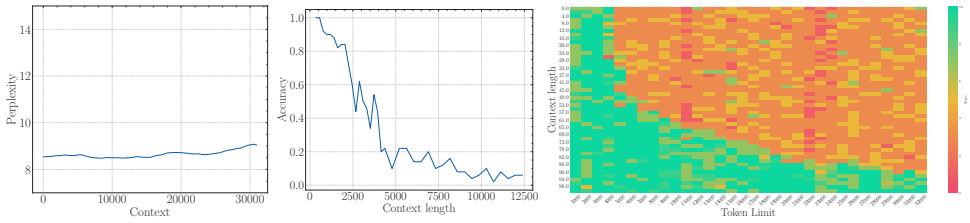


Figure 12: Baichuan2-7B-Base with base=1e4 fine-tuned on 32k context (original context=4096)

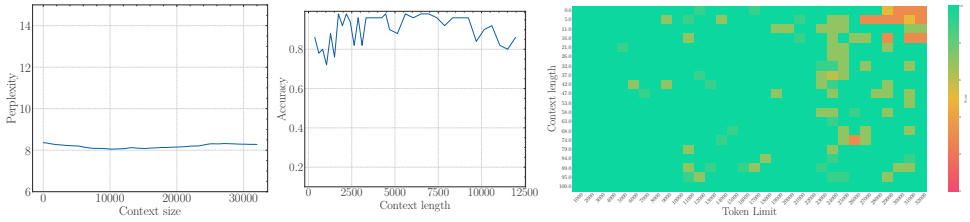


Figure 13: Baichuan2-7B-Base with base=2e5 fine-tuned on 32k context (original context=4096)

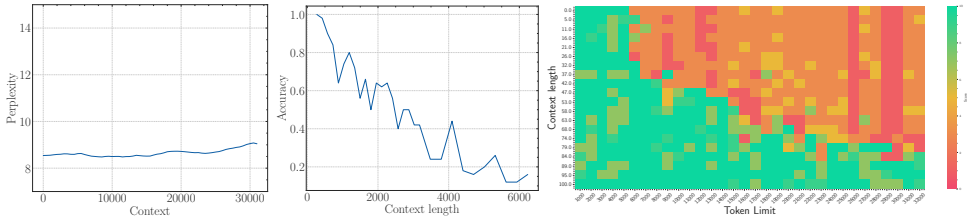


Figure 14: Qwen1.5-7B-Base [9] with base=1e4 fine-tuned on 32k context (original context=4096)

E The python code for calculating the low bound base for a context length of 32k

```python
"""the python code for calculate the
low bound base for a context length of 32k"""
import torch
import numpy as np
def get_BMtheta_expectation(base,context_size=2**15,dim=128):
    realdim = dim / 2
    d= torch.arange(0, realdim, 1)
    theta = base ** (-2*d/dim)
    dist= torch.outer(torch.arange(0,context_size),theta).cos()
    return dist.sum(dim=1) / realdim
search_base = []
for x in range(3,10):
    for i in range(1,10):
        for j in range(10):
            search_base.append((i+j/10)* (10**x))
for base in search_base:
    ans = get_BMtheta_expectation(base)
    if True not in (ans<0):
        print("Find!Base=", base)
        break
    idx = np.argmax(ans < 0)
    print('base', base, 'first zero position', idx)
```

F A empirical verification of Desiderata 2

The **Desiderata 2** introduced in Section 4 is intuitively plausible, but its empirical validity requires verification. To investigate this, we conducted a detailed empirical analysis. The similarity between tokens is measured by the cosine similarity (denoted as A) of their corresponding hidden states, while the attention allocation between tokens is governed by the attention score (denoted as B). The desiderata "similar tokens receive more attention" implies that a higher value of **A** should lead to a higher value of **B**.

To test this desiderata, we performed experiments using Llama1-7B, Llama2-7B, and Llama3-8B models. We selected 200 segments from the PG19 dataset, each containing 1024 tokens, and computed Spearmans rank correlation coefficient between (A) and (B). A positive correlation coefficient would indicate that as token similarity (A) increases, the corresponding attention score (B) also increases. The magnitude of the coefficient reflects the strength of this correlation.

The results, presented in Figure 15 confirm that Spearmans rank correlation coefficient is positive, validating the desiderata that "similar tokens receive more attention". Furthermore, we observe that this positive correlation is more pronounced in the Llama3-8B model compared to Llama2-7B and Llama1-7B, suggesting that larger and more advanced models are better at capturing this relationship.

G Evaluation results on RULER

https://doi.org/10.52202/079017-2773

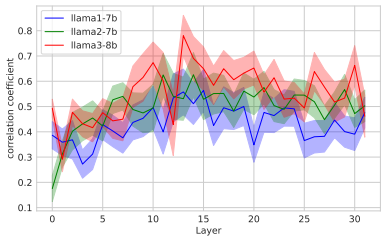


Figure 15: Spearmans rank correlation coefficient between the similarity and the attention score of two tokens. The thick line represents the mean μ calculated from different samples. The upper and lower boundaries of the line are $\mu + \sigma$ and $\mu - \sigma$, respectively, where σ is the standard deviation of different samples.

Table 5: Evaluation results on RULER. We finetune Llama2-7b to 32k context length (the low bound base is 6e5) using different RoPE's bases. NS is short for NIAH-single and NM is short for NIAH-Multikey.

Base	Context Len.	Sub tasks													Ave.
		NS-1	NS-2	NS-3	NM-1	NM-2	NM-3	NIAH_Multivalue	NIAH_Multiquery	VT	CWE	FWE	QA1	QA2	
500	4k	23.0	31.0	26.0	31.0	13.0	11.0	73.0	75.0	2.0	59.5	53.7	51.0	29.0	36.78
	8k.	15.0	18.0	11.0	14.0	2.0	3.0	50.0	37.0	1.0	46.3	46.0	18.0	21.0	21.72
	16k	5.0	8.0	5.0	9.0	2.0	1.0	28.0	33.0	0.0	23.0	40.7	23.0	27.0	15.74
	32k	1.0	1.0	2.0	4.0	1.0	1.0	10.0	12.0	0.0	1.5	22.7	16.0	24.0	7.40
1e4	4k	99.0	100	96.0	91.0	85.0	65.0	66.0	99.0	90.0	34.1	77.33	66.0	44.0	77.88
	8k.	53.0	55.0	58.0	59.0	34.0	4.0	49.0	84.0	1.0	33.7	27.67	30.0	29.0	39.80
	16k	21.0	24.0	28.0	36.0	17.0	3.0	72.0	75.0	0.0	49.3	8.67	10.0	25.0	28.38
	32k	5.0	8.0	11.0	13.0	7.0	0.0	38.0	39.0	0.0	17.1	1.33	19.0	26.0	14.19
2e5	4k	100	100	100	97.0	97.0	77.0	99.0	99.0	100	79.6	86.0	45.0	45.0	86.51
	8k.	100	100	100	100	96.0	48.0	97.0	100	100	42.9	65.00	44.0	40.0	79.46
	16k	100	100	100	97.0	74.0	23.0	92.0	100	97.0	20.7	8.33	38.0	37.0	68.23
	32k	99.0	100.0	95.0	95.0	32.0	9.0	62.0	87.0	82.0	27.0	39.0	29.0	38.0	61.08
6e5	4k	100	100	100	97.0	96.0	65.0	99.0	100	100	84.6	90.0	52.0	49.0	87.12
	8k.	100	100	100	99.0	96.0	40.0	93.0	100	100	43.4	66.33	34.0	47.0	78.36
	16k	100	100	100	95.0	74.0	37.0	93.0	99.0	98.0	27.4	62.67	37.0	41.0	74.16
	32k	100	100	94.0	96.0	47.0	12.0	70.0	89.0	97.0	20.5	63.67	25.0	39.0	65.63
9e5	4k	100	100	99.7	97.0	95.1	71.0	99.0	99.7	100	83.6	88.5	49.3	46.9	86.91
	8k.	100	100	100	98.4	96.3	48.4	92.7	100	100	44.66	67.53	35.8	46.7	79.27
	16k	100	100	100	93.8	78.8	42.4	90.9	99.3	98.6	27.58	59.97	36.4	40.1	74.45
	32k	100	100	95.8	96.3	52.7	18.3	64.3	89.6	97.9	17.26	63.77	26.2	39.0	66.24
5e6	4k	100	100	99.0	97.0	93.0	85.0	99.0	99.0	100.0	81.2	85.0	43.0	42.0	86.40
	8k	100	100	100	97.0	97.0	68.0	92.0	100	100	47.6	70.3	40.0	46.0	81.38
	16k	100	100	100	100	91.0	90.0	55.0	86.0	100	100	28.0	53.7	35.0	79.90
	32k	100	100	100	97.0	66.0	33.0	51.0	91.0	100.0	9.7	64.0	29.0	39.0	67.67
1e9	4k	100	100	100	95.0	96.0	72.0	100	99.0	67.0	63.8	77.7	41.9	29.0	80.11
	8k	100	100	100	96.0	90.0	54.0	95.0	100	88.0	35.0	60.0	28.0	35.0	75.46
	16k	100	100	100	96.0	77.0	43.0	83.0	100	72.0	23.7	51.3	27.0	35.0	69.85
	32k	100	100	100	93.0	69.0	23.0	58.0	92.0	94.0	18.1	55.7	17.0	35.0	65.75

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: We claims the paper's contributions and scope in the abstract and the last part of the introduction in Section 1.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: We discuss the limitations of the work in Section 7.

   Guidelines:

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

3. **Theory Assumptions and Proofs**

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

   Answer:[Yes]

https://doi.org/10.52202/079017-2773

Justification: We provide the assumptions of Theorem 1 in itself and provide the proof in Appendix A.

Guidelines:

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

4. **Experimental Result Reproducibility**

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

   Answer: [Yes]

   Justification: We provide the experimental details and hype-parameters in Section 5.1 and Appendix B.

   Guidelines:

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

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

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

Answer: [No]

Justification: Due to some privacy policies and the complexity of training large language models, we are unable to provide the data and code. But we believe that based on the open-source LLMs and open-source code, our results can be reproduced.

Guidelines:

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

6. **Experimental Setting/Details**

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

Answer: [Yes]

Justification: We provide the experimental details and hype-parameters in Section 5.1 and Appendix B.

Guidelines:

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

7. **Experiment Statistical Significance**

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

Answer: [No]

Justification: Due to the high computational cost of large language models, it is difficult for us to perform the same experiment multiple times to get the error bar. However, we provide the results of different models and their performance under various evaluation metrics to support our perspective.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

 https://doi.org/10.52202/079017-2773

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

8. **Experiments Compute Resources**

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

   Answer: [Yes]

   Justification: We provide the information in Section B.

   Guidelines:

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

9. **Code Of Ethics**

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

   Answer: [Yes]

   Justification: We have read the NeurIPS Code of Ethics and the research conducted in our paper conform, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

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

10. **Broader Impacts**

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

    Answer: [NA]

    Justification: Our paper focuses on the impact of key parameter settings in the model on its capability. To our knowledge, this does not involve any social impact.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.

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

11. **Safeguards**

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

    Answer: [NA]

    Justification: In this work, we don't release any data or model.

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer:[Yes]

    Justification: We cite existing papers and url.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

https://doi.org/10.52202/079017-2773

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

13. **New Assets**

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

    Answer: [NA]

    Justification: The paper does not release new assets.

    Guidelines:

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

14. **Crowdsourcing and Research with Human Subjects**

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

    Answer: [NA]

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

    Guidelines:

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

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

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

    Answer: [NA]

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

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

https://doi.org/10.52202/079017-2773