On the token distance modeling ability of higher RoPE attention dimension

Anonymous EMNLP submission

Abstract

Length extrapolation algorithms based on Rotary position embedding (RoPE) have shown promising results in extending the context length of language models. However, understanding how position embedding can capture longer-range contextual information remains elusive. Based on the intuition that different dimensions correspond to different frequency of changes in RoPE encoding, we conducted a dimension-level analysis to investigate the correlation between a hidden dimension of an attention head and its contribution to capturing long-distance dependencies. Using our correlation metric, we identified a particular type of attention heads, which we named *Positional Heads*, from various length-extrapolated models. These heads exhibit a strong focus on long-range information interaction and play a pivotal role in long input processing, as evidence by our ablation. We further demonstrate the correlation between the efficiency of length extrapolation and the extension of the high-dimensional attention allocation of these heads. The identification of Positional Heads provides insights for future research in long-text comprehension.

1 Introduction

The Transformer model has revolutionized natural language processing tasks, but it demonstrates limitations in modeling long sequences. Meanwhile, models like Mamba (Gu and Dao, 2023) that excel in capturing long-range dependencies struggle to meet the practical requirements of natural language modeling (Lieber et al., 2024). Consequently, there has been a recent surge of work focused on extending the context length in language models based on the Transformer architecture (Zhang et al., 2024; Xiong et al., 2023; Fu et al., 2024b). Particularly, some of these efforts that leverage and enhance the capabilities of RoPE (Rotary Positional Embedding) (Jin et al., 2024; Peng et al., 2023; Chen et al., 2023a), have shown promising results in extrapolating the model's capacity to handle longer contexts (Wang et al., 2024).

Open-source large language models commonly employ Rotary Positional Embedding (RoPE) to model sequence positional information (Touvron et al., 2023; Jiang et al., 2023; Yang et al., 2023; Bai et al., 2023a). RoPE exhibits two desirable properties. Firstly, its exponential positional encoding introduces long-range attention decay, allowing the model to focus more on neighboring semantic information. Secondly, by utilizing trigonometric functions to differentiate frequencies, RoPE effectively captures different distances between tokens, enabling higher attention scores for tokens that have longer semantic dependencies, facilitating semantic aggregation. When compared to length extrapolation methods based on sparse attention (Ratner et al., 2022; Xiao et al., 2023) or prompt compression (Yen et al., 2024; Xiao et al., 2024b), modifications to RoPE for length extrapolation do not result in the loss of fine-grained contextual information at a global level. Therefore, it possesses distinct advantages in tasks such as long text comprehension (Bai et al., 2023b; Lv et al., 2024), where the preservation of comprehensive contextual information is essential for practical applications.

A prevailing viewpoint suggests that language models based on RoPE encounter out-of-distribution (OOD) issues when faced with contexts longer than the pre-training text length, specifically affecting the sampling of the trigonometric function component for token distances (Peng et al., 2023; Xiong et al., 2023). As a result, related studies have adjusted the attention resolution in the context of long texts and fine-tuned the model to adapt to longer token distances. We hypothesize that the effectiveness of such methods stems from RoPE's ability to decouple information from different distances by representing them through different dimensions with varying rotational frequencies.

However, this pattern has not been thoroughly observed and analyzed in the actual inference process.

Our paper presents a novel approach by examining the impact of each dimension of RoPE, specifically focusing on the 128-dimensional attention heads, on modeling text distances. We empirically validate the claim in Yarn that lower-frequency dimensions are responsible for modeling longer text dependencies. Furthermore, we discover that not all attention heads exhibit this characteristic, emphasizing the importance of heads that possess this relationship in modeling long texts. Our study explores RoPE's potential for long text modeling from a frequency perspective, shedding light on the relationship between dimensions and text modeling capabilities. Our primary findings are as follows: nosep

- In most attention heads, regardless of whether length extrapolation is performed, the impact of high-dimensional low-frequency components is greater than that of low-dimensional high-frequency components.

- Input lengths exceeding the pre-training length can result in anomalies in high-dimensional components. Length extrapolation extend the high-dimensional attention allocation for longer token distance.

- We refer attention heads that have stronger correlation between token distance and dimension allocation as *Positional Heads*, which play a crucial role in modeling text distances.

2 Background

2.1 Rotary Position Embeddings

Large Language Models (LLMs) are primarily based on the Transformer architecture (Vaswani et al., 2017), with the attention mechanism at its core. A prevalent method for incorporating positional information in these models is Rotary Position Embeddings (RoPE) (Su et al., 2021), which leverages rotation matrices to encode the positional information of sequences.

In RoPE, the positional encoding for a hidden layer, with the hidden dimension denoted by d, uses a rotation matrix for each position m. The rotation matrix $\mathcal{R}_m$ is defined as follows:

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

where

$$\theta_i = 10000^{-2i/d} \tag{2}$$

Explicitly, for the query vector $\mathbf{q}$ at position m and the key vector $\mathbf{k}$ at position n, we have:

$$\mathbf{q} = \begin{bmatrix} q_0 \\ q_1 \\ \vdots \\ q_{d-1} \end{bmatrix}, \quad \mathbf{k} = \begin{bmatrix} k_0 \\ k_1 \\ \vdots \\ k_{d-1} \end{bmatrix} \tag{3}$$

After applying RoPE, the transformed vectors $\mathbf{q}_m$ and $\mathbf{k}_n$ are given by:

$$\mathbf{q}_m = \mathcal{R}_m \mathbf{q} = \begin{bmatrix} q_{m0} \\ q_{m1} \\ \vdots \\ q_{m(d-1)} \end{bmatrix}, \mathbf{k}_n = \mathcal{R}_n \mathbf{k} = \begin{bmatrix} k_{n0} \\ k_{n1} \\ \vdots \\ k_{n(d-1)} \end{bmatrix} \tag{4}$$

The attention weights are then calculated using the dot product of the transformed vectors:

$$\text{softmax}\left(\frac{\mathbf{q}_m^T \mathbf{k}_n}{\sqrt{d}}\right) \tag{5}$$

The dot product for $\mathbf{q}_m$ and $\mathbf{k}_n$ is given by:

$$\mathbf{q}_m^T \mathbf{k}_n = \sum_{i=0}^{d-1} q_{m,i} k_{n,i} \tag{6}$$

2.2 Length Extrapolation Methods

We have investigated methods to extend the context length of language models, particularly using Rotary Position Embedding (RoPE). Our research focuses on three prominent techniques: Yarn (Peng et al., 2023), CLEX (Chen et al., 2023a), and Self-Extend (Jin et al., 2024). Each method leverages different aspects of positional encoding to enhance long-range token interactions, showing favorable performance in our tests.

YaRN (Peng et al., 2023) addresses Out-of-Distribution (OOD) scenarios by categorizing RoPE dimensions into three frequency-based groups and applying tailored interpolation strategies. Low-frequency dimensions use linear interpolation with adjusted θ_i (=1) for smooth transitions. High-frequency dimensions remain unchanged, while intermediate-frequency dimensions use linear interpolation to bridge the extremes effectively.

CLEX (Chen et al., 2023a) advances the concept of Dynamic Scaling by modeling $\theta_i(\text{pos})$ as a continuous function of position using a neural ODE. This method enables precise parameter fine-tuning to fit the data, demonstrating superior performance in our tests.

SelfExtend (Jin et al., 2024) uses bi-level attention: grouped attention and neighbor attention, to capture dependencies among both distant and adjacent tokens. It addresses positional O.O.D. issues by remapping unseen large relative positions to those encountered during pretraining through a floor division operation. This approach allows LLMs to maintain coherence over longer texts without fine-tuning.

3 Defining Dimension Contribution in RoPE

In Rotary Position Embedding (RoPE), each dimension of the vectors $\mathbf{q}_m$ and $\mathbf{k}_n$ contributes to the attention score via their dot product. To thoroughly investigate the role of different dimensions in RoPE for semantic modeling, we utilize an algorithm that analyzes the contribution of each dimension to the attention scores.

To capture the contribution of each dimension, we employ the Hadamard product, i.e., element-wise multiplication, denoted by the symbol $\odot$:

$$\mathbf{h} = \mathbf{q}_m \odot \mathbf{k}_n \in R^d, \mathbf{h}_i = q_{m,i}k_{n,i}, \qquad (7)$$

where

$$\begin{aligned} h_{2i} &= q_{2i}k_{2i}\cos(m\theta_i)\cos(n\theta_i) \\ &\quad - q_{2i+1}k_{2i}\sin(m\theta_i)\cos(n\theta_i) \\ &\quad - q_{2i}k_{2i+1}\cos(m\theta_i)\sin(n\theta_i) \\ &\quad + q_{2i+1}k_{2i+1}\sin(m\theta_i)\sin(n\theta_i) \\ h_{2i+1} &= q_{2i}k_{2i}\sin(m\theta_i)\sin(n\theta_i) \\ &\quad + q_{2i+1}k_{2i}\cos(m\theta_i)\sin(n\theta_i) \\ &\quad + q_{2i}k_{2i+1}\sin(m\theta_i)\cos(n\theta_i) \\ &\quad + q_{2i+1}k_{2i+1}\cos(m\theta_i)\cos(n\theta_i) \qquad (8) \end{aligned}$$

In RoPE, every two dimensions correspond to trigonometric functions with the same frequency θ_i. We sum the values of these corresponding dimensions to form new vectors:

$$\mathbf{g} \in R^{\frac{d}{2}}, \quad g_i = h_{2i} + h_{2i+1} \qquad (9)$$

for $i = 0, 1, 2, \ldots, \frac{d}{2} - 1$.

The value of g_i reflects the contribution of each dimension in RoPE to the attention score. A higher value indicates a greater contribution of that dimension to the attention score, where:

$$\begin{aligned} g_i &= h_{2i} + h_{2i+1} \\ &= (q_{2i}k_{2i} + q_{2i+1}k_{2i+1})\cos((m-n)\theta_i) \\ &\quad + (q_{2i}k_{2i+1} - q_{2i+1}k_{2i})\sin((m-n)\theta_i). \end{aligned} \qquad (10)$$

Here, θ_i represents the positional encoding frequency for the i-th dimension.

The dot product of $\mathbf{q}_m$ and $\mathbf{k}_n$ can be expressed as:

$$\mathbf{q}_m^T\mathbf{k}_n = \sum_{i=0}^{d-1} q_{m,i}k_{n,i} = \sum_{i=0}^{d-1} h_i = \sum_{i=0}^{\frac{d}{2}-1} g_i \qquad (11)$$

Therefore, we use the value g_i to measure the contribution of θ_i to the attention score.

This methodological framework enables a comprehensive analysis of how each dimension in Rotary Position Embedding (RoPE) contributes to the attention scores.Through this approach, we can delve into the role of different dimensions in RoPE for semantic modeling.

4 Experiments

4.1 Study on dimension-level contributions to attention scores

This study aims to answer the following question: ***Are there distinct patterns of attention contributions across different dimensions?***

To examine this, we initially observe the overall contribution of each dimension to the attention scores. We sampled 17 inputs, and for each input, at each layer and each head of the model, we randomly selected 100 × number of tokens qk pairs. For each selected qk pair, we computed the contribution of each dim as shown in Section 3, then recorded the top 5 dimensions that contributed the most. We conducted a statistical analysis of the distribution of attention scores in terms of dimensions for each layer and head across four models: the original Llama-2-7B, Mistral-7B, and their vrsions with 64K length extrapolation using the Yarn method. The average values of these dimension distributions are presented in Figure 1. As per the theoretical analysis of RoPE, the models tend to focus on syntactic parsing in the shallow layers, placing greater emphasis on shorter distance information. The attention scores of the majority of attention heads are predominantly contributed by the higher-dimensional components. There were no significant changes observed in the dimension distribution before and after length extrapolation.

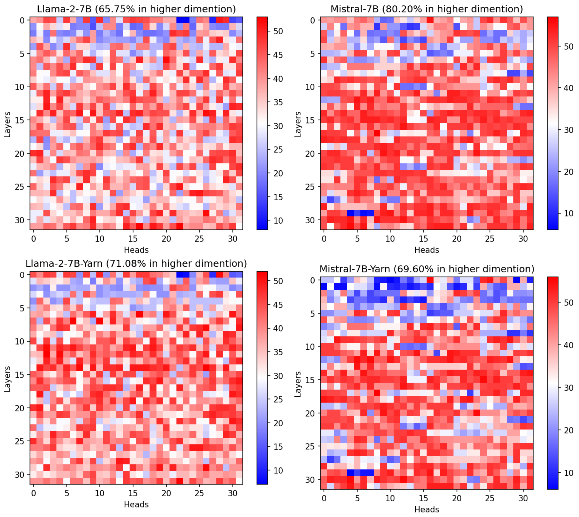

Figure 1: The average of the dimensional distribution of attention scores for each head in each layer of the four models

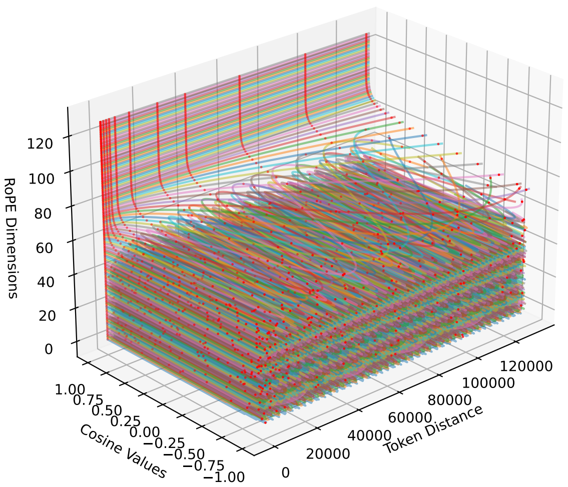

Figure 2: The value of trigonometric function in rotation position coding changes with the dimension and token distance, and the red dots represents the trigonometric function value of different dimension corresponding to a specific token distance

We now provide a potential explanation for the higher contribution observed in the higher-dimensional components. According to Equation (11), the effect of the rotated positional encoding matrix on positions n and m in dimension i is equivalent to a trigonometric function of the form $cos(n-m)\theta_i$. We visualized the values of these trigonometric functions, as shown in Figure 2. The red dots represent several distances (n-m) with the maximum token distance set at 128k. It can be observed that tokens with longer distances correspond to shorter distinguishable curves in the rotated positional encoding. In the lower-dimensional range, the abrupt changes in values between adjacent dimensions become irregular due to the higher frequency of the trigonometric function. The purpose of RoPE is to encode different token distances in different dimensions, and the aggregation of information from different dimensions is performed when computing attention scores. The irregularity in the lower-dimensional range hinders the disentanglement of distance-related information. Consequently, during the training process, the model tends to favor the working of attention in the higher-dimensional components. Moreover, the maximum text length that the model can handle is also determined by the higher-dimensional components. Furthermore, increasing the base of the exponential function lowers the frequency of the trigonometric function, leading to increased distinguishable components in the higher dimensions. This has been confirmed to be a practical method for length extrapolation in pre-training approaches such as Llama3(AI@Meta, 2024) and Code-Llama (Roziere et al., 2023).

4.2 Study on correlation between dimensions and token distances

The previous study confirms the significant contribution from higher dimensions. Continued from the conclusion, in this study, we aim to understand: ***Are higher dimensions responsible for long-range attention among tokens?***

4.2.1 Correlation Plot

Indeed, according to the principles of RoPE, higher dimensions are responsible for modeling longer token distances. However, it remains to be examined whether this correlation strictly holds in the actual inference process of pre-trained models such as Llama. In order to investigate this, we primarily focused on Llama and employed the methods showed below to observe the original Llama model as well as three different length extrapolation meth-

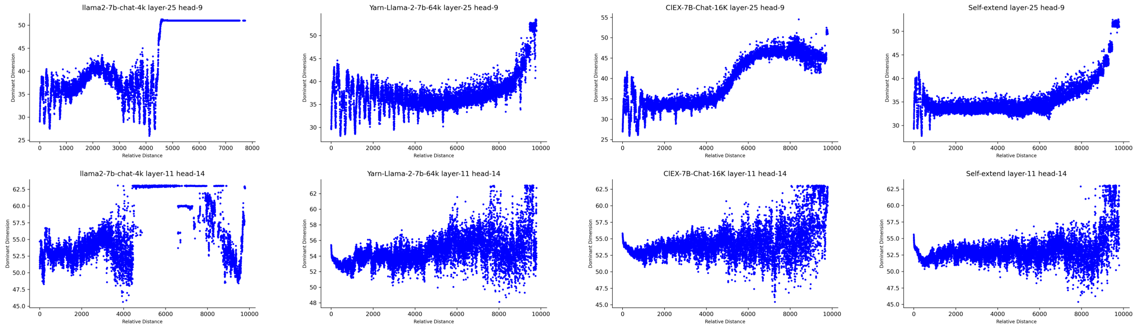

Figure 3: Correlation plot comparing the original Llama model with three different length extrapolation methods. The top and bottom rows show correlations in different heads.

ods. To comprehensively assess the influence of all dimensions in Rotary Position Embedding (RoPE) on a given query-key pair ($\mathbf{q_m}$ and $\mathbf{k_n}$), we propose an algorithm to compute the *Dominant Dimension*. This value is determined by analyzing the contribution scores assigned to each dimension within RoPE. The *Dominant Dimension* signifies that the attention score predominantly originates from the vicinity of this particular dimension. For each vector $\mathbf{g}_i$ in (9), we apply the softmax function:

$$\text{softmax}(\mathbf{g})_i = \frac{e^{\mathbf{g}_i}}{\sum_j e^{\mathbf{g}_j}} \tag{12}$$

We then compute the dot product of the softmax output with its corresponding position vector to determine the *Dominant Dimension*:

$$\textit{Dominant Dimension} = \text{softmax}(\mathbf{g}) \cdot \mathbf{pos} \tag{13}$$

where

$$\mathbf{pos} = \begin{bmatrix} 0 & 1 & \ldots & \frac{d}{2} - 1 \end{bmatrix}^T \tag{14}$$

To investigate the relationship between relative distance and Dominant Dimension, we sampled 17 prompts. For each prompt, across every layer and head of the model, we selected the top 100 tokens with the highest interaction attention scores for each token. This resulted in a total of 100 times the number of tokens qk pairs. For each qk pair, *Dominant Dimension* was computed, and its relative distance $m - n$ was recorded.

For each head, we obtained a collection of 100× number of tokens *Relative Distance - Dominant Dimension* pairs. If a relative distance corresponds to multiple Dominant Dimensions, we averaged them to obtain the Dominant Dimension corresponding to that distance.

The correlation between token relative distances and the dominant dimension of attention scores is depicted in Figure 3.

4.2.2 Observation

Through a thorough analysis of the relationship between the dominant dimension and the relative distance of each head of each layer of the model, we have drawn the following inspiring observation:

1. In some heads of the model, there is a significant correlation between the dominant dimension and the relative distance, whereas, in other heads, this correlation is not observed.
2. For the original Llama model, a sudden change in the dominant dimension occurs when the sequence length exceeds the pretraining length (4K). We observed a similar phenomenon in other models, such as Baichuan, as illustrated in Appendix A.1.
3. For the length extrapolation method, by observing the dominant dimension of the model, it can be seen that this method extends the trend of the dominant dimension within the pre-training length range of Llama to a new length range, thereby achieving length extrapolation. This observation is consistent with the design methodology of these length extrapolation approaches.

4.2.3 OOD Explanation

To further elucidate why the original model exhibits a sudden change in behavior when exceeding the pretraining length, we conducted an ablation study on the dimension matrix R_m of the rotary position encoding. The results are depicted in the figure. As shown in Figure 4, it can be observed that when the length is less than the pretraining length (4K), the image after removing R_m shows little difference compared to the original. However, beyond the pretraining length, no abrupt changes occur. Therefore, we propose a plausible explanation based on

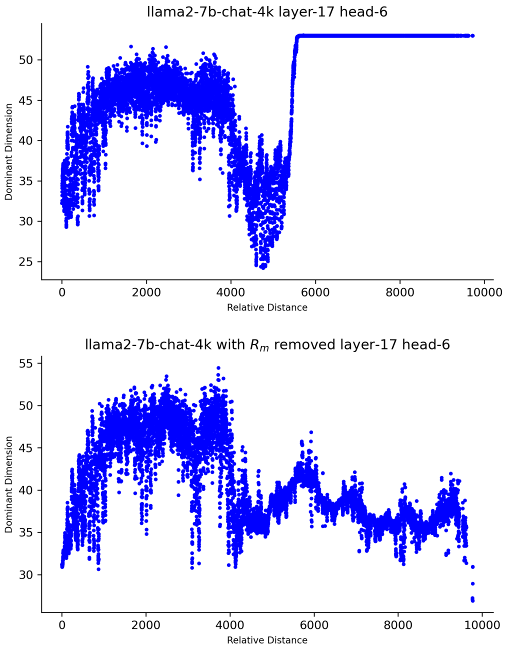

Figure 4: Correlation plot comparing the original Llama model with the model where R_m has been removed. On the left side are the results from the original Llama model, and on the right side are the results after removing R_m. It is evident that the abrupt changes disappear after R_m removal.

the finding: As depicted in Figure 2, as the relative distance increases, the lower dimensions of the rotary positional encoding tend to resemble characteristics similar to random sampling, while the higher dimensions remain comparatively stable. Consequently, when the relative position exceeds the pretraining length (4K), the values in the lower dimensions gradually become overshadowed by noise from the trigonometric functions, whereas the values in the higher dimensions remain intact. The model training adjusts to accommodate this sampling characteristic of trigonometric functions. However, when the relative distance surpasses the model's pretraining length, the model struggles to adapt to this extended sampling range, leading to a scenario where the lower dimensions lose coherence while the influence of the higher dimensions becomes predominant.

4.3 Finding the *Positional Heads*

4.3.1 Positional Heads Detection

Positional Heads refer to attention heads with significant correlations mentioned in Section 4.2.2. In order to identify them, we quantified the distance-dimension correlation for each head using the Spearman rank correlation coefficient. This statistical measure was computed based on the visualization provided earlier. A Spearman correlation coefficient closer to 1 (in absolute value) indicates a stronger correlation, with the sign showing the direction. More details are in Appendix A.2.

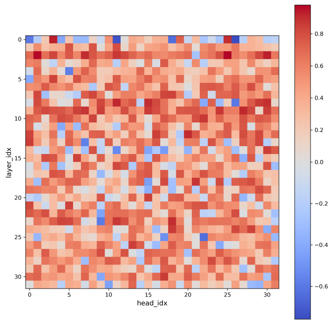

Figure 5: Spearman correlation coefficients of each head in the YaRN-Llama-2-7b-64K model. In most heads, there is a correlation between the dominant dimension and the relative distance.

4.3.2 Influence of Positional Heads on long distance modeling

To validate the importance of attention heads with high distance-dimension correlations for long text comprehension, we conducted a masking procedure on these heads. Using the metrics described in the previous section, we identified the top 5% and top 10% heads based on their rankings and set their output to zero. We then compared the performance of these heads with randomly sampled 5% and 10% heads. The results, as shown in follows, demonstrate that heads with high distance-dimension correlations exhibit greater importance across various tasks.

Question Ansering. The question-answering (QA) task is a commonly used text comprehension task that requires models to comprehend long text inputs and retrieve information relevant to the given questions. We utilized four QA tasks from LongBench (Bai et al., 2023b) to evaluate the impact of random masking of attention heads on the results. Figure 6 indicate that randomly masking out attention heads had no significant effect on the results. However, when we masked out the top 5% and 10% heads based on the distance-dimension correlation metrics, it resulted in a significant decline in the model's performance on this task. More results can be found in Appendix A.3.

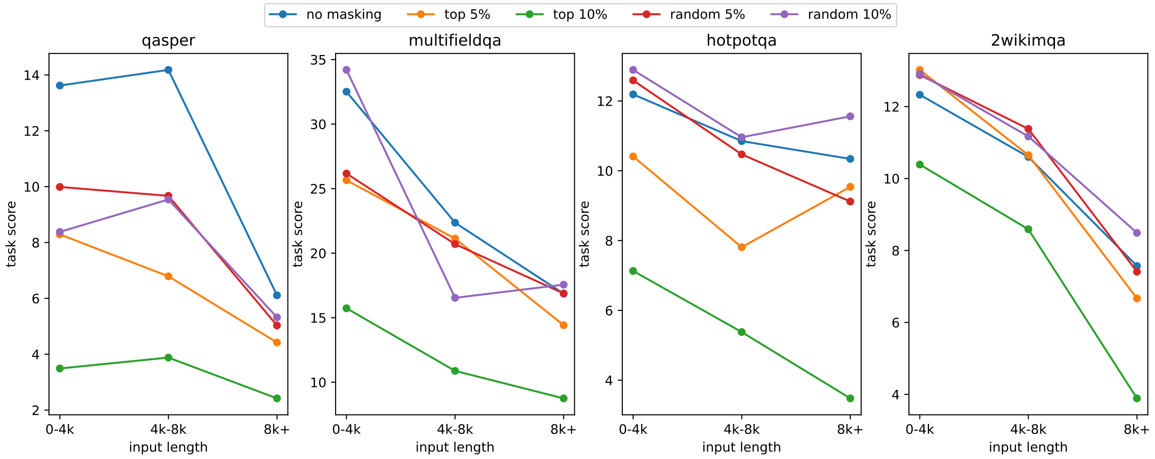

Figure 6: Masking out top scored heads v.s. random heads. For the QA tasks in LongBench, the removal of heads with top scores clearly reduces performance.

Code Completion. Compared to the QA task, the code completion task places higher demands on long-distance dependencies in the text. We employed the code completion task from LongBench to assess the impact of random masking of attention heads on the results. As shown in Figure 7, this observation suggests that our proposed metrics can effectively identify the heads that are more important for understanding long texts from the perspective of long-distance information interaction.

PassKey. The PassKey task is commonly used to evaluate the long text retrieval capability of models. We conducted the same ablation experiments on this task. The models used were the original Llama2-7B model and the Llama2-7B model with length extrapolation using the Yarn method. The results are shown in Figure 8. When the input length exceeds the pre-training length of the model, the original model exhibits out-of-distribution failures in long-distance retrieval. However, when we mask out the high-score attention heads of the length-extrapolated model, the model shows a uniform performance decline across all lengths of retrieval, indicating that these attention heads are highly sensitive to the distance between texts. On the other hand, random masking out of attention heads does not exhibit this phenomenon.

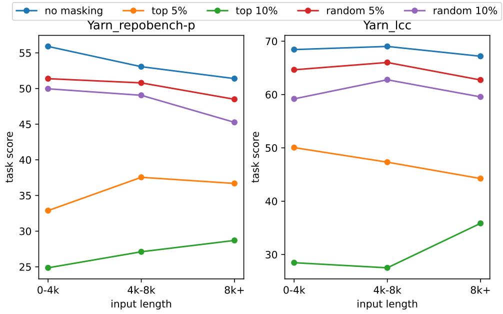

Figure 7: Masking out top scored heads v.s. random heads. For the Code tasks in LongBench, the removal of heads with top scores clearly reduces performance.

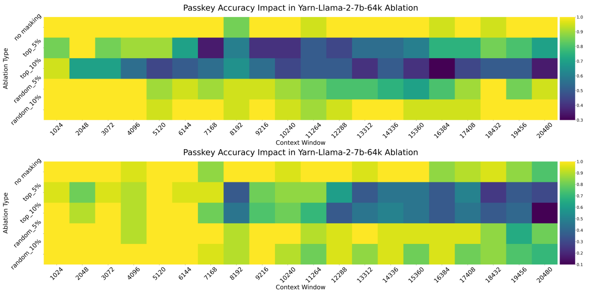

Figure 8: Masking out heads with top scores v.s. random heads. For the passkey task , the removal of heads with top scores clearly reduces performance.

Perplexity. While evaluating the long text comprehension ability of the models, it is important to ensure that the fundamental performance of the models does not collapse. We assessed the perplexity (PPL) of the aforementioned models and their ablated versions, and the results are shown in Figure 9. It can be observed that although the ablation of high-score attention heads led to a decrease in PPL, it did not result in the PPL explosionseen in the original Llama model when faced with long texts.

5 Related Work

Handling longer contexts in Transformer models has seen significant improvements through various

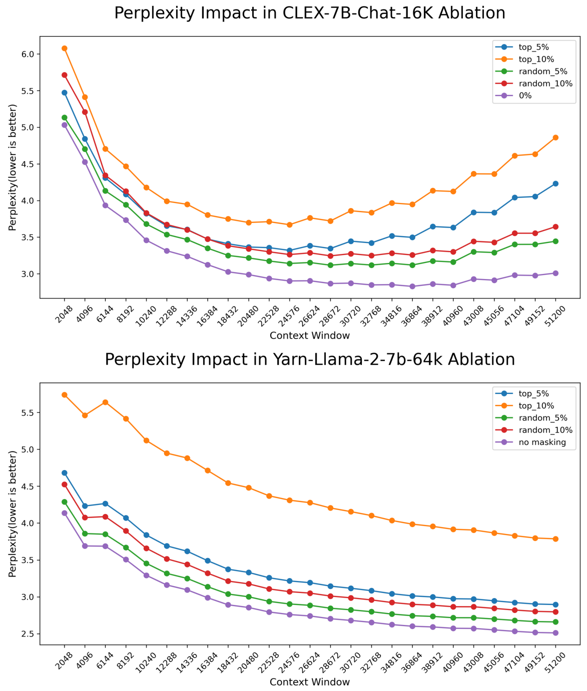

Figure 9: Masking out heads with top scores v.s. random heads. For the ppl, the removal of heads with top scores clearly reduces performance.

methods, including enhanced training techniques, innovative frameworks, memory mechanisms, and adjustments to positional encoding, including enhanced training techniques(Fu et al., 2024a), external summary designs (Xiao et al., 2024a), memory mechanisms (Dai et al., 2019; Mohtashami and Jaggi, 2023), and adjustments to positional encoding. Among these methods, modifying positional encoding stands out due to its simplicity and efficacy. Some methods manipulate the token position numbering itself, as seen in PI (Chen et al., 2023b) and Selfextend (Jin et al., 2024). Others make adjustments within the encoding layers at the level of rotational positional encoding, exemplified by works like YaRN (Peng et al., 2023) and CLEX (Chen et al., 2023a). Additionally, novel positional encodings, such as CoPE (Golovneva et al., 2024), have been proposed to generalize rotational positional encoding and further enhance long-text capabilities.

Certain studies have delved into the impact of positional encoding in depth. Some research indicates that the initial token's position is crucial in long-text contexts (Han et al., 2023; Xiao et al., 2023), while other work (Men et al., 2024) highlights that the base of rotational positional encoding can limit a model's capacity to handle long texts. Fang et al. (2024) proposes a comprehensive framework to describe length extrapolation. However, the role of different dimensions within rotational positional encoding for information interaction remains underexplored. Moreover, the precise mechanisms by which positional encoding affects information interaction are not yet fully understood.

Highly relevant to this work are studies focusing on the interpretability of attention heads(Wu et al., 2024; Olsson et al., 2022). These studies specifically investigate the role of attention heads in information retrieval processes. The function of self-attention mechanisms extends beyond mere replication of highly relevant information; we emphasize the capability of self-attention mechanisms to integrate information from different positions. This capability is crucial for practical long text comprehension tasks.

6 Conclusion

We investigated the properties of attention heads with rotary position embeddings (RoPE) in commonly used Transformer architectures. Using long text comprehension tasks as a starting point, we explored the modeling of token-to-token distance within the model by deconstructing the contributions of different dimensions within the attention heads to the attention scores.

We found that due to the computational nature of rotary position embeddings, higher dimensions of the attention heads, which correspond to lower rotational frequencies, are more effective at distinguishing distances between tokens. Furthermore, attention heads that, through training, allocate attention scores across different dimensions according to token distances and exhibit a certain degree of correlation, demonstrate superior capabilities in modeling text distances. These heads are crucial for integrating information from varying distances in long text comprehension tasks.

We provide an analytical perspective on the currently popular rotary position embeddings, illustrating the attention patterns of models trained with RoPE. Future research can leverage the properties of these attention heads to address challenging tasks such as long text comprehension.

Limitations

Although we demonstrated the capability of RoPE in modeling textual distances, several limitations are worth noting. First, our dimensional decomposition approach is based on the explicit meaning of dimensions in rotary position embeddings; this method is not applicable to all types of position encodings. Nonetheless, we maintain that decoupling token distance in attention computation is crucial for integrating and understanding information across different distances. Second, due to computational resource constraints, we could not implement many hypotheses we wished to validate on a larger scale. Our observations were not validated with longer input sequences, and the impact of fine-tuning on these attention heads was not analyzed. We leave more detailed experimental analysis to future work.

References

AI@Meta. 2024. Llama 3 model card.

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

Yushi Bai, Xin Lv, Jiajie Zhang, Hong Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2023b. Longbench: A bilingual, multitask benchmark for long context understanding. *ArXiv*, abs/2308.14508.

Guanzheng Chen, Xin Li, Zaiqiao Meng, Shangsong Liang, and Li Bing. 2023a. Clex: Continuous length extrapolation for large language models. *ArXiv*, abs/2310.16450.

Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. 2023b. Extending context window of large language models via positional interpolation. *ArXiv*, abs/2306.15595.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime G. Carbonell, Quoc V. Le, and Ruslan Salakhutdinov. 2019. Transformer-xl: Attentive language models beyond a fixed-length context. In *Annual Meeting of the Association for Computational Linguistics*.

Junjie Fang, Likai Tang, Hongzhe Bi, Yujia Qin, Si Sun, Zhenyu Li, Haolun Li, Yongjian Li, Xin Cong, Yukun Yan, Xiaodong Shi, Sen Song, Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2024. Unimem: Towards a unified view of long-context large language models. *ArXiv*, abs/2402.03009.

Yao Fu, Rameswar Panda, Xinyao Niu, Xiang Yue, Hanna Hajishirzi, Yoon Kim, and Hao Peng. 2024a. Data engineering for scaling language models to 128k context. *ArXiv*, abs/2402.10171.

Yao Fu, Rameswar Panda, Xinyao Niu, Xiang Yue, Hannaneh Hajishirzi, Yoon Kim, and Hao Peng. 2024b. Data engineering for scaling language models to 128k context. *arXiv preprint arXiv:2402.10171*.

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

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

Chi Han, Qifan Wang, Wenhan Xiong, Yu Chen, Heng Ji, and Sinong Wang. 2023. Lm-infinite: Simple on-the-fly length generalization for large language models. *ArXiv*, abs/2308.16137.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

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

Opher Lieber, Barak Lenz, Hofit Bata, Gal Cohen, Jhonathan Osin, Itay Dalmedigos, Erez Safahi, Shaked Meirom, Yonatan Belinkov, Shai Shalev-Shwartz, et al. 2024. Jamba: A hybrid transformer-mamba language model. *arXiv preprint arXiv:2403.19887*.

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

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

Amirkeivan Mohtashami and Martin Jaggi. 2023. Landmark attention: Random-access infinite context length for transformers. *ArXiv*, abs/2305.16300.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova Dassarma, Tom Henighan, Benjamin Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, John Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom B. Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Christopher Olah. 2022. In-context learning and induction heads. *ArXiv*, abs/2209.11895.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. 2023. Yarn: Efficient context window extension of large language models. *ArXiv*, abs/2309.00071.

Nir Ratner, Yoav Levine, Yonatan Belinkov, Ori Ram, Inbal Magar, Omri Abend, Ehud Karpas, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2022. Parallel context windows for large language models. *arXiv preprint arXiv:2212.10947*.

Baptiste Roziere, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Tal Remez, Jérémy Rapin, et al. 2023. Code llama: Open foundation models for code. *arXiv preprint arXiv:2308.12950*.

Jianlin Su, Yu Lu, Shengfeng Pan, Bo Wen, and Yunfeng Liu. 2021. Roformer: Enhanced transformer with rotary position embedding. *ArXiv*, abs/2104.09864.

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. 2023. Llama: Open and efficient foundation language models. *ArXiv*, abs/2302.13971.

Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Neural Information Processing Systems*.

Xindi Wang, Mahsa Salmani, Parsa Omidi, Xiangyu Ren, Mehdi Rezagholizadeh, and Armaghan Eshaghi. 2024. Beyond the limits: A survey of techniques to extend the context length in large language models. *arXiv preprint arXiv:2402.02244*.

Wenhao Wu, Yizhong Wang, Guangxuan Xiao, Hao Peng, and Yao Fu. 2024. Retrieval head mechanistically explains long-context factuality. *ArXiv*, abs/2404.15574.

Chaojun Xiao, Pengle Zhang, Xu Han, Guangxuan Xiao, Yankai Lin, Zhengyan Zhang, Zhiyuan Liu, Song Han, and Maosong Sun. 2024a. Infllm: Training-free long-context extrapolation for llms with an efficient context memory.

Chaojun Xiao, Pengle Zhang, Xu Han, Guangxuan Xiao, Yankai Lin, Zhengyan Zhang, Zhiyuan Liu, Song Han, and Maosong Sun. 2024b. Infllm: Unveiling the intrinsic capacity of llms for understanding extremely long sequences with training-free memory. *arXiv preprint arXiv:2402.04617*.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. 2023. Efficient streaming language models with attention sinks. *ArXiv*, abs/2309.17453.

Wenhan Xiong, Jingyu Liu, Igor Molybog, Hejia Zhang, Prajjwal Bhargava, Rui Hou, Louis Martin, Rashi Rungta, Karthik Abinav Sankararaman, Barlas Oguz, et al. 2023. Effective long-context scaling of foundation models. *arXiv preprint arXiv:2309.16039*.

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

Howard Yen, Tianyu Gao, and Danqi Chen. 2024. Long-context language modeling with parallel context encoding. *arXiv preprint arXiv:2402.16617*.

Peitian Zhang, Zheng Liu, Shitao Xiao, Ninglu Shao, Qiwei Ye, and Zhicheng Dou. 2024. Soaring from 4k to 400k: Extending llm's context with activation beacon. *arXiv preprint arXiv:2401.03462*.

A Appendix

A.1 OOD phenomenon in other model

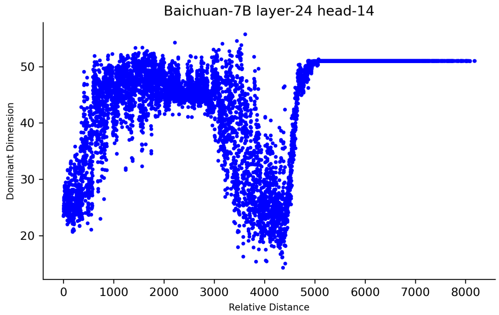

Figure 10: Correlation plot of Baichuan-7B. When the sequence length exceeds the pre-training length of 4k, the dominant dimension of Baichuan exhibits a sudden change.

A.2 Spearman Correlation Coefficient in Different Methods

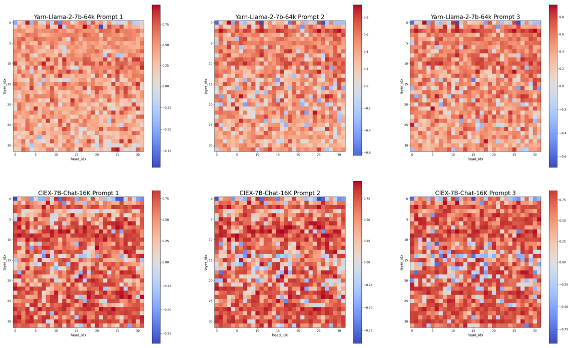

Figure 11: Spearman correlation coefficients of each head in the YaRN-Llama-2-7b-64K model and CLEX-7B-chat-16K across different prompts, illustrating the stability of position heads across prompts. In most heads, there is a notable correlation between the dominant dimension and the relative distance.

A.3 Other masking Results

	Qasper			MultifieldQA			HotpotQA			2WikiMQA		
masking method	0-4k	4-8k	8k+	0-4k	4-8k	8k+	0-4k	4-8k	8k+	0-4k	4-8k	8k+
Selfextend-no-masking	19.52	16.27	21.39	40.73	34.77	27.25	45.5	41.86	40.21	40.12	32.64	28.07
Selfextend-Random5%	14.05	16.07	5.33	37.83	24.57	23.47	42.8	39.73	36.97	39.49	33.14	22.11
Selfextend-Random10%	15.19	15.08	3.95	37.57	23.81	18.97	44.03	39.42	31.71	33.94	29.63	20.45
Selfextend-Top5%	17.43	12.27	4.59	37.74	23.52	20.62	42.53	16.17	7.91	29.19	17.85	6.46
Selfextend-Top10%	8.19	7.39	3.75	31.92	17.89	14.28	33.39	14.57	5.54	30.7	12.12	5.15
CLEX-no-masking	25.06	27.69	19.94	48.31	32.88	24.75	21.42	23.88	28.0	21.76	20.55	9.01
CLEX-Random5%	21.52	26.12	15.88	43.43	31.52	24.0	24.22	17.87	22.22	19.92	18.58	11.02
CLEX-Random10%	22.55	27.15	19.7	46.94	30.56	19.75	25.85	24.65	29.59	19.0	18.5	14.29
CLEX-Top5%	17.59	22.34	12.04	42.57	31.13	34.73	23.22	24.61	27.17	21.77	18.36	14.44
CLEX-Top10%	13.64	18.82	8.66	45.34	29.56	22.63	21.49	22.18	24.26	17.99	20.05	12.83

Table 1: Self-extend performance on QA tasks when masking out heads with top scores vs. random heads. Removing heads with top scores significantly reduces performance.