On the token distance modeling ability of higher RoPE attention dimension

Xiangyu Hong[1*], Che Jiang[1*], Biqing Qi[1†], Fandong Meng[2], Mo Yu[2] , Bowen Zhou[1†] , Jie Zhou[2]

[1]Department of Electronic Engineering, Tsinghua University, [2]Pattern Recognition Center, WeChat AI, Tencent Inc, China

Correspondence to: hong-xy22@mails.tsinghua.edu.cn, jc23@ mails.tsinghua.edu.cn, zhoubowen@tsinghua.edu.cn

Introduction:

- Length extrapolation algorithms based on Rotary position embedding (RoPE) have shown promising results in extending the context length of language models.
- Our study investigates how RoPE can capture longer-range contextual information.
- Our primary findings are as follows :
  - **1. High-dimensional components of RoPE with low rotary frequency have a greater impact than low-dimensional high-frequency components**.
  - **2. Exceeding pre-training input lengths causes high-dimensional anomalies, while length extrapolation methods extend high-dimensional attention allocation over longer distances.**
  - **3. Attention heads with strong token-distance and dimension correlation, called Positional Heads, are key for modeling text distances.**

Background and Definition

RoPE(Rotary Position Embedding)

query $\mathbf{q}$ at position m key $\mathbf{k}$ at position n

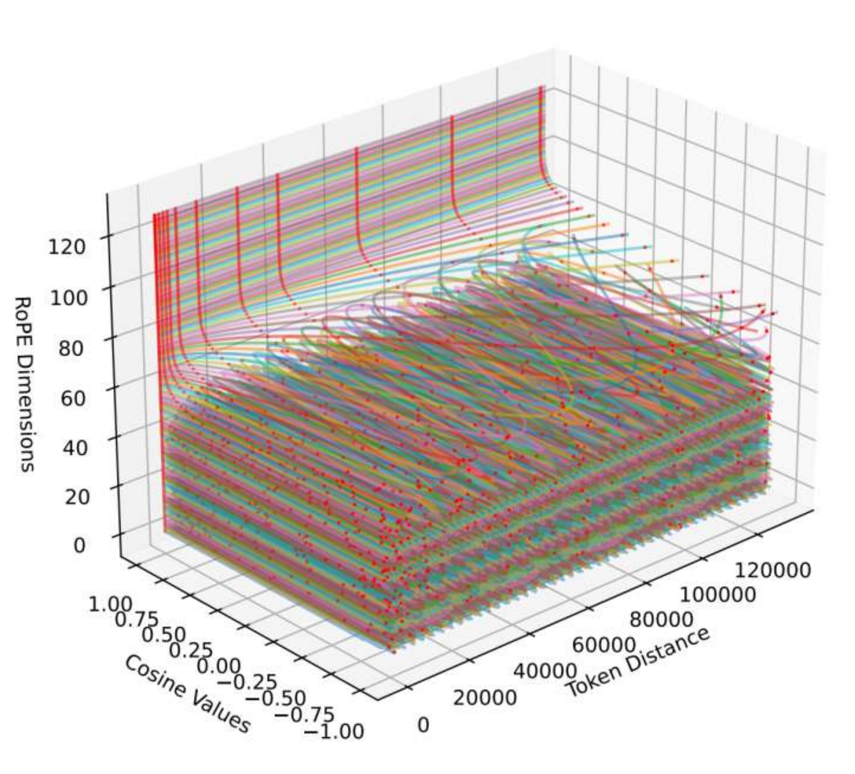
$$\mathbf{q}=\begin{bmatrix}q_0\\q_1\\\vdots\\q_{d-1}\end{bmatrix}\quad \mathbf{k}=\begin{bmatrix}k_0\\k_1\\\vdots\\k_{d-1}\end{bmatrix}\quad \underbrace{\begin{pmatrix}\cos m\theta_0 & -\sin m\theta_0 & 0 & 0 & \cdots & 0 & 0\\ \sin m\theta_0 & \cos m\theta_0 & 0 & 0 & \cdots & 0 & 0\\ 0 & 0 & \cos m\theta_1 & -\sin m\theta_1 & \cdots & 0 & 0\\ 0 & 0 & \sin m\theta_1 & \cos m\theta_1 & \cdots & 0 & 0\\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots\\ 0 & 0 & 0 & 0 & \cdots & \cos m\theta_{d/2-1} & -\sin m\theta_{d/2-1}\\ 0 & 0 & 0 & 0 & \cdots & \sin m\theta_{d/2-1} & \cos m\theta_{d/2-1}\end{pmatrix}}_{\mathcal{R}_m}$$

$$\mathbf{q}_m=\mathcal{R}_m\mathbf{q}\qquad \mathbf{k}_n=\mathcal{R}_n\mathbf{k}$$

$$\text{attn weight}=(\mathcal{R}_m\boldsymbol{q})^\top(\mathcal{R}_n\boldsymbol{k}) = \boldsymbol{q}^\top\mathcal{R}_m^\top\mathcal{R}_n\boldsymbol{k}=\boldsymbol{q}^\top\mathcal{R}_{n-m}\boldsymbol{k}\qquad \theta_i=10000^{-2i/d}$$

m : postion of the token

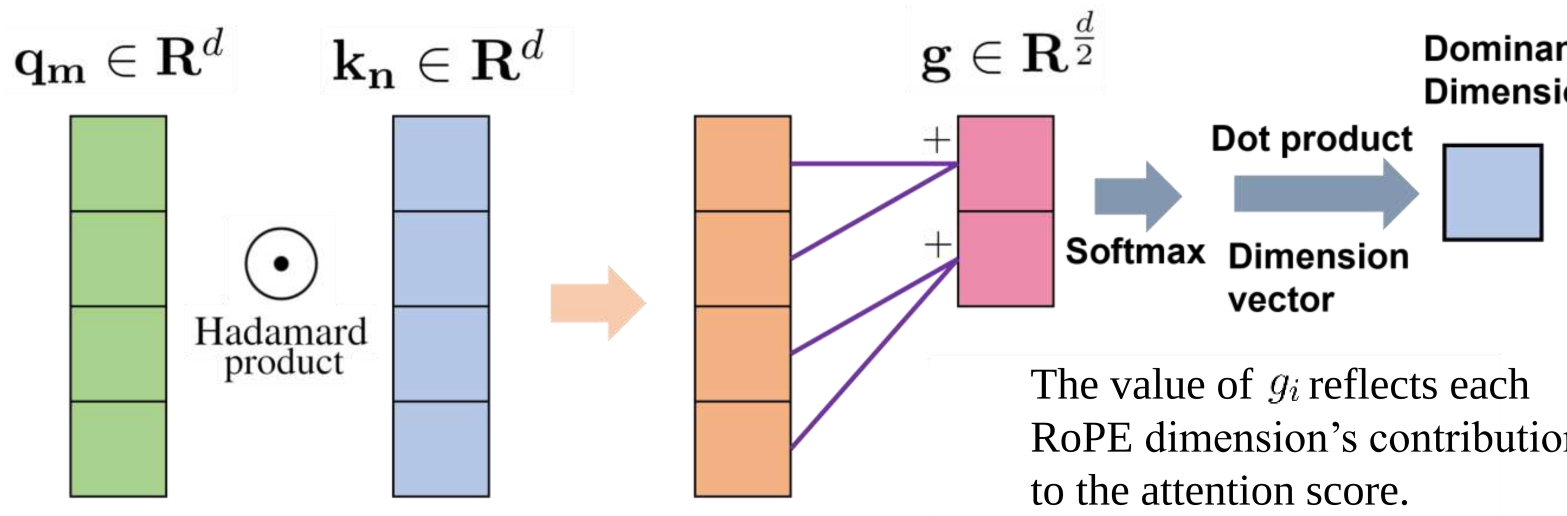
Properties of RoPE



- **Colored lines:** The value of trigonometric function in rotation position coding changes with the dimension and token distance
- **Red dots:** the trigonometric function value of different dimension corresponding to some specific token distance

Figure corresponds to the $\mathcal{R}_{n-m}$ matrix
- Token Distance axis : n-m
- RoPE Dimensions axis : i in θ_i
- Cosine Values axis: $\cos(n-m)\theta_i$

Defining dimension contribution



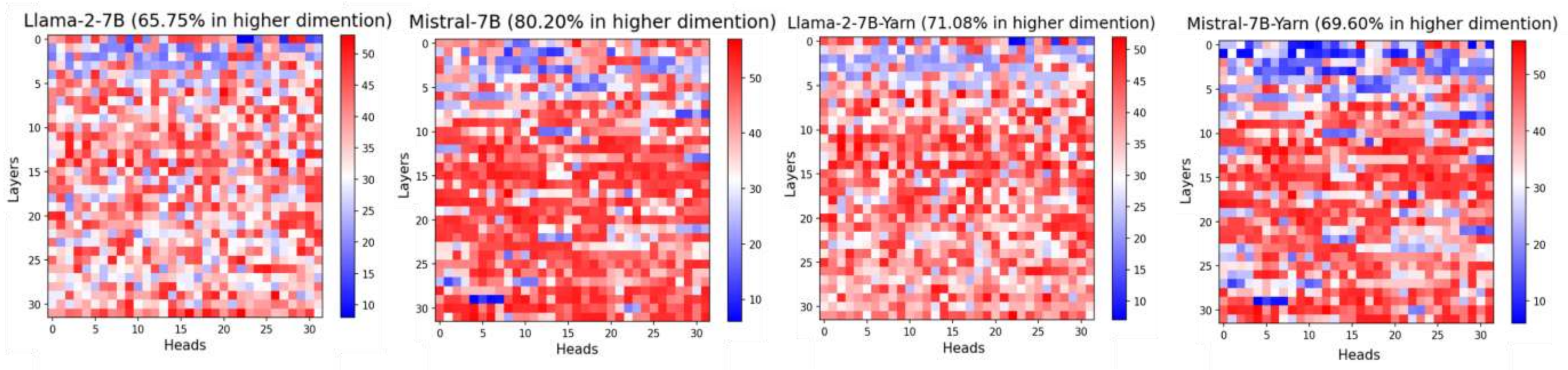
The value of g_i reflects each RoPE dimension's contribution to the attention score.

Experiment Setup

- Dataset: LongBench
- Model: Llama2-7B-chat (We use greedy decoding strategy for consistency)

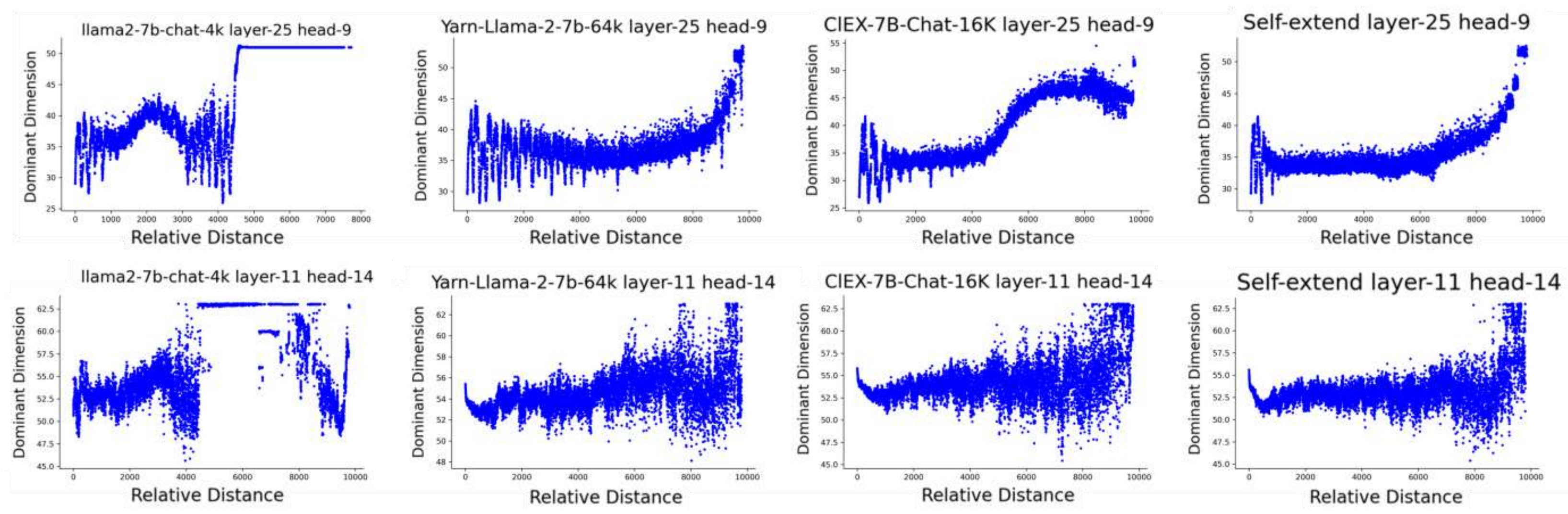
Results:

1. Are there distinct patterns of attention contributions across different dimensions?

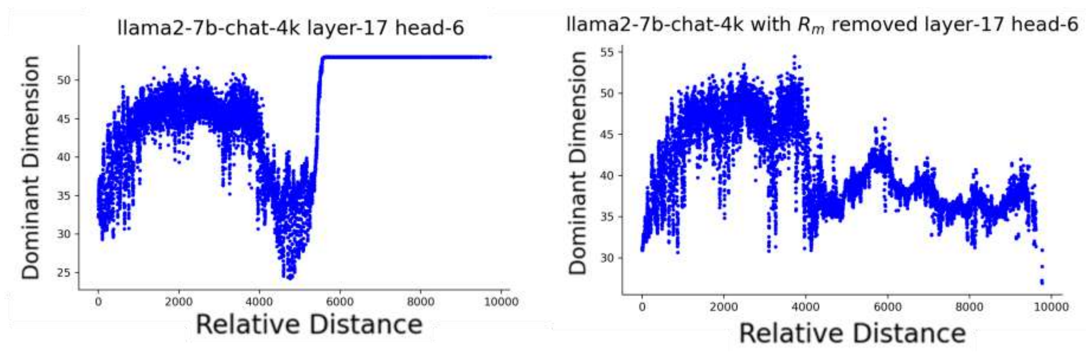


The average dimensional distribution of attention scores for each head in every layer across the four models shows that **higher dimensions of RoPE contribute more to the attention weights**.

2. Are higher dimensions responsible for long-range attention among tokens?

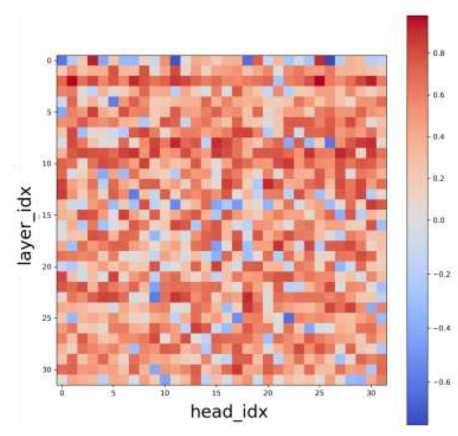


- There is a significant relationship between the **dominant dimension** and **relative distance** in some heads of the model (like L25H9), while this correlation is absent in others (like L11H14).
- Length extrapolation methods extend the original correlation pattern (column 1) to a new length range (column 2-4), which aligns with the design methodology of these approaches.



Abrupt changes when exceeding the pretraining length disappear after removing RoPE matrix, indicating that R_m is responsible for token distance OOD.

3. Finding and Ablating the Positional Heads



Spearman correlation coefficients of each head in the YaRN-Llama-2-7b-64K model.