Massive Values in Self-Attention Modules are the Key to Contextual Knowledge Understanding

Mingyu Jin [1] **Kai Mei** [1] **Wujiang Xu** [1] **Mingjie Sun** [2]
Ruixiang Tang [1] **Mengnan Du** [3] **Zirui Liu** [4] † **Yongfeng Zhang** [1] †

ICML 2025

Background

- Our understanding of LLMs' internal mechanisms and how these mechanisms relate to observable behaviors remains limited

- Existing research about LLM representations:
  - Residual stream activations can exhibit massive values
  - Massive values appear exclusively in Q and K while are absent in V
  - Massive values have been identified as critical factors influencing quantization
  - Existing studies do not explore the rationale behind this phenomenon deeply

- **This paper systematically investigates the formation of massive values and their connection to model behaviors**

Key Findings

- Massive values are concentrated in **specific regions** of Q and K exclusively.
  - These massive values in each head's dim index are very close
  - Absent in V and absent in models without RoPE

- Massive values in Q and K are **critical for understanding contextual knowledge** over parametric knowledge
  - Disrupting these values leads to a notable degradation in tasks requiring contextual understanding

- Quantization techniques **targeting massive values** preserve contextual knowledge better

- Concentration of massive values is **caused by RoPE** and it **appears since very early layers** in LLMs

Definition of Massive Value

What Is Massive Value?

- Attention queries (Q) and keys (K) are represented as:

Batch Size=1 | Seq Len | Number of Heads | Head Dim

$$Q, K \in \mathbb{R}^{\mathcal{B} \times \mathcal{S} \times \mathcal{H} \times \mathcal{D}}$$

Seq: Only consider the input prompt

- Compute the L2 norm along the sequence length dimension:

$$M_{h,d} = \|Q_{:,h,d}\|_2 = \sqrt{\sum_{s=1}^{\mathcal{S}} Q_{s,h,d}^2}.$$

- Massive Value:

$$M_{h,d} > \lambda \frac{1}{\mathcal{D}} \sum_{d'=1}^{\mathcal{D}} M_{h,d'}$$

$\lambda > 1$ (=5 in the paper) is a threshold controlling massive value selection

Concentrate Massive Value

In each attention head, **certain dimensions** exhibit notably massive values, and these tend to cluster **in specific dimensional regions**.

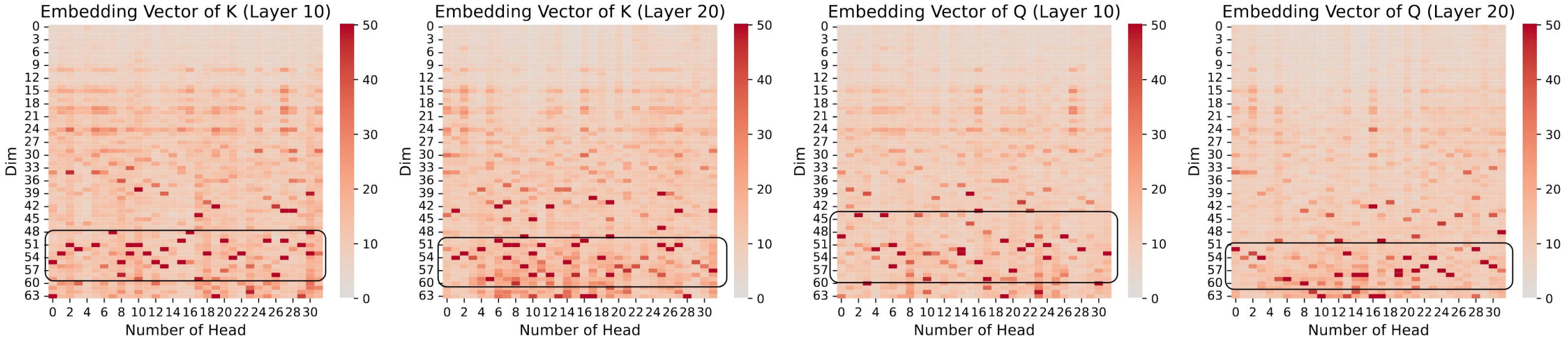


Figure 2. Q and K Embedding Vector in Llama-2-7B, we choose Layer 10 and 20, and the input question is shown as Figure 11. This visualization shown here is a two-dimensional image because we averaged over the sequence-length dimension. The horizontal axis is the *number of head* and the vertical axis is *head dim*. We can see that the massive value is concentrated at the bottom of the picture.

Massive Values Play A Key Role in Contextual Knowledge Understanding

Tasks

Contextual Knowledge Understanding vs. Parametric Knowledge Retrieval

Math	GSM8k	1k
Math	AQUA	1k
Movie Review	IMDB	1k
Synthetic Passkey Retrieval		没找到

An example of Passkey Retrieval Task (6-128)

Prompt: There is important info hidden inside a lot of irrelevant text. Find it and memorize it. I will quiz you about the important information there. The grass is green. The sky is blue. The sun is yellow. Here we go. There and back again. The grass is green. The sky is blue. The sun is yellow. Here we go. There and back again. The grass is green. The sky is blue. The sun is yellow. Here we go. There and back again. The pass key is 383816. Remember it. 383816 is the passkey.
Ground Truth: 383816

Factual QA	Cities	1k
Synthetic Factual QA(True/False)		800

Table 5. Examples of parameter knowledge retrieval task: factual QA, covering Sports, Arts, Technology and Celebrity.

Category	Example	Ground Truth
Sports	Is the Olympic Games held every four years?	Yes
	Was Babe Ruth a famous football player?	No
	Is the FIFA World Cup held every two years?	No
Arts	Was the painting 'Girl with a Pearl Earring' completed during the 18th century?	No
	Is Pablo Picasso one of the founding figures of Cubism?	Yes
	Was Diego Rivera a famous Mexican muralist?	Yes
Technology	Is the ASCII character set limited to 256 characters?	No
	Was the first iPhone released in 2007?	Yes
	Is Linux an open-source operating system?	Yes
Celebrity	Is Leonardo DiCaprio an Oscar-winning actor?	Yes
	Was Taylor Swift born in Los Angeles?	No
	Was Michael Jackson a member of The Beatles?	No

Disruption of Massive Values and Non-Massive Values

Let $X \in R^{lxhxd}$ denote the query tensor. *l是prompt长度*

Disruption of Massive Values: Replace the values (both Q and K) at the massive value indices with the

$$\mathbf{X}_{i;j;k^*} = \begin{cases} \text{Mean}(\mathbf{X}) , & k^* = \text{argmax}_k x_{i;j;k} \\ \mathbf{X}_{i;j;k^*} , & k^* \neq \text{argmax}_k x_{i;j;k} \end{cases}$$

Disruption of Non-massive Values: Replace the top n (from 1 to 20) smallest values (both Q and K) with calculated averages.

$$\mathbf{X}_{i;j;k^*} = \begin{cases} \text{Mean}(\mathbf{X}), & k^* = \arg\min x_{i;j;k} \\ \mathbf{X}_{i;j;k^*}, & k^* \neq \arg\min x_{i;j;k} \end{cases}$$

Massive Values Contribute to Contextual Knowledge Understanding

Table 1. Results of LLMs under different settings (vanilla, massive value disrupted, non-massive value disrupted) on different benchmarks. For the Passkey Retrieval Task, the values (max prompt token length, passkey length) represent the maximum number of tokens allowed in the prompt and the length of the passkey to be retrieved, respectively. All values are reported in percentage (%).

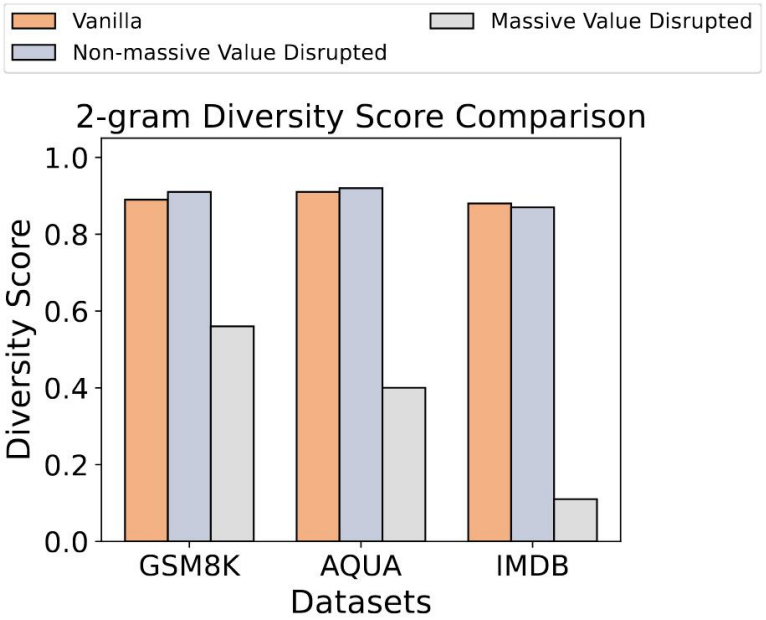
Model	Contextual Knowledge Understanding Task						Parametric Knowledge Retrieval Task				
	GSM8K	AQUA	Passkey Retrieval Task			IMDB	Cities	Sports	Art	Technology	Celebrity
			(128,6)	(256,12)	(1024,48)						
Gemma2-9B	81.30	63.80	100	100	100	94.70	99.70	91.00	84.00	81.00	92.50
+ Non-Massive Value Disrupted	81.60	65.60	100	100	100	97.40	99.60	91.00	84.00	81.50	92.50
+ Massive Value Disrupted	15.10	16.50	2.00	0.00	0.00	1.80	76.40	73.50	68.00	72.00	82.00
Llama3-8B	76.90	53.51	100	100	100	95.40	99.40	95.00	93.50	92.50	95.00
+ Non-Massive Value Disrupted	77.40	53.90	100	100	100	95.40	99.40	94.50	93.00	92.50	95.50
+ Massive Value Disrupted	4.00	9.68	9.00	0.00	0.00	11.00	88.20	74.50	64.00	74.90	73.00
Qwen2.5-7B	86.60	56.69	100	100	100	96.80	97.70	95.00	96.00	90.00	93.50
+ Non-Massive Value Disrupted	85.40	57.28	100	100	100	97.60	97.50	94.00	96.50	90.00	93.50
+ Massive Value Disrupted	16.10	19.68	9.00	1.00	0.00	6.53	81.50	74.00	69.50	71.00	71.00

When only non-massive values are disrupted, performance remains remarkably stable across all tasks and models

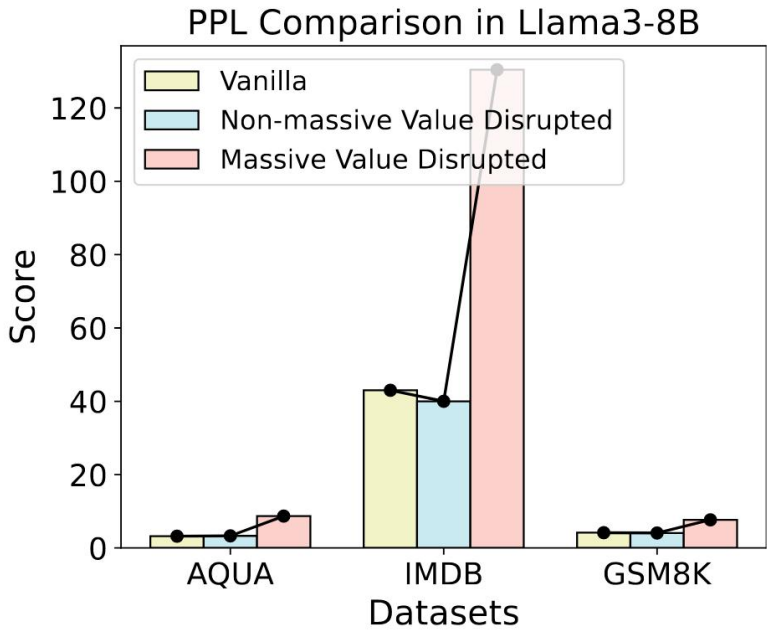
Parametric Knowledge Retrieval tasks maintain relatively high accuracy even when massive values are disrupted

Contextual Knowledge Understanding Tasks drop a lot when massive values are disrupted

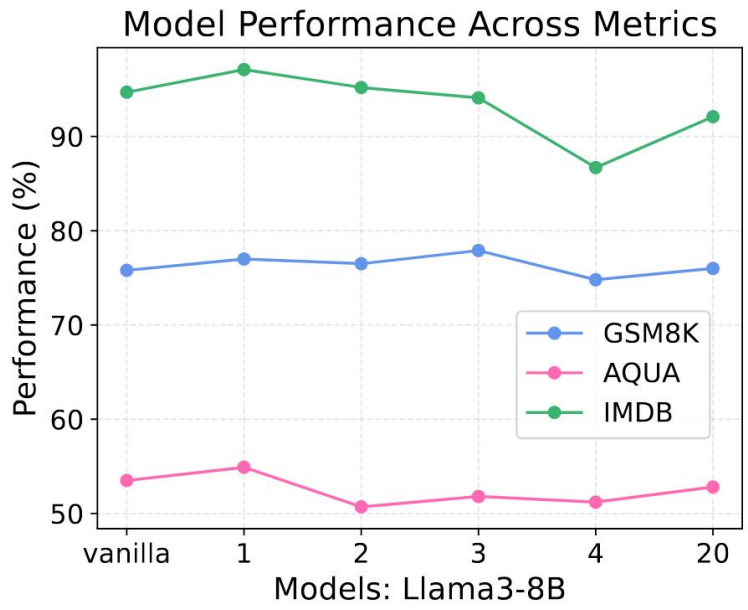
Massive Values Contribute to Contextual Knowledge Understanding



(a) The 2-gram diversity scores for Llama3-8B under three conditions across three datasets.



(b) The perplexity scores of the Llama3-8B model under three conditions across three datasets.



(c) The figure shows Llama3-8B maintains relatively performance under different non-massive value disrupted methods.

Figure 3. Disrupting massive values leads to higher perplexity and lower diversity, while disrupting non-massive values maintains model performance, particularly evident in IMDB dataset analysis.

Effects of Massive Values on Knowledge Conflict and Quantization

When massive values are destroyed, the model **loses its ability to process misleading contextual information** and instead defaults to its parametric knowledge.

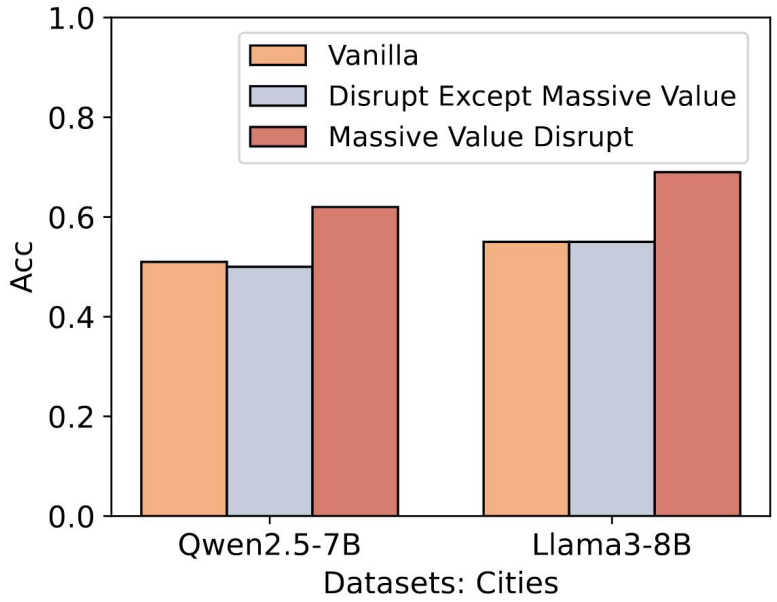


Figure 4. We can observe that introducing conflicting background knowledge causes LLM to be misled into making random guesses. However, **after massive values are disrupted, the model is still able to maintain a certain level of accuracy.**

Quantization methods that **protect massive values** maintain good and robust performance on contextual knowledge understanding tasks.

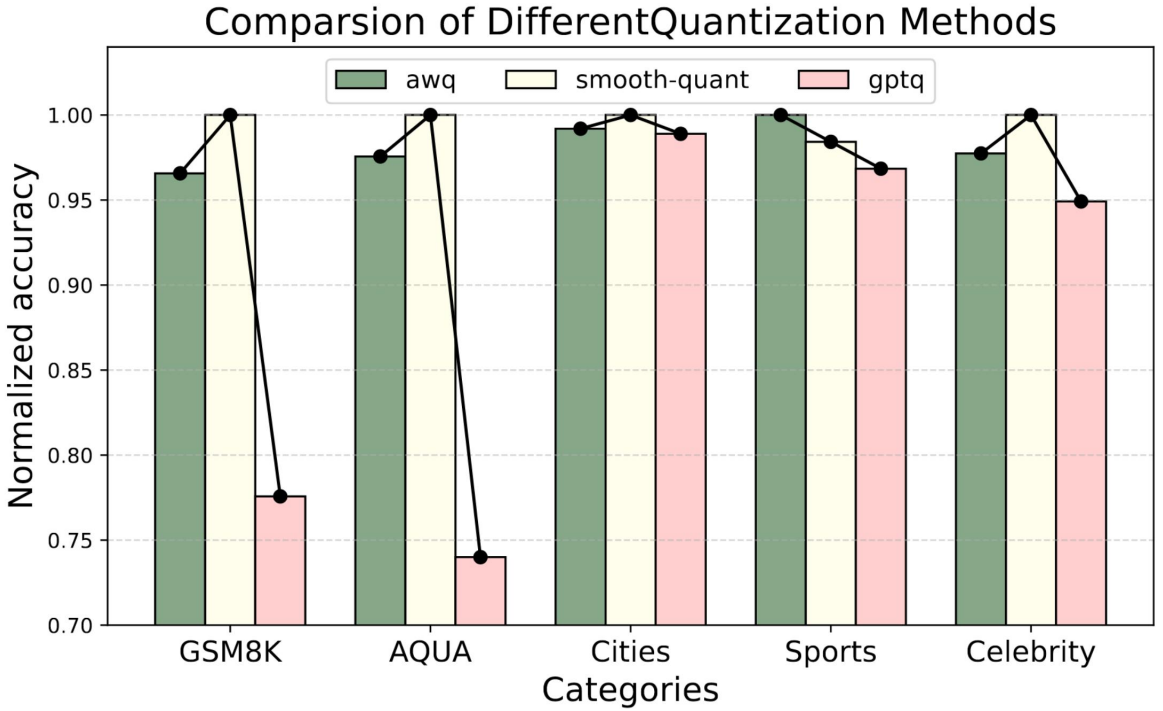


Figure 5. Impacts of different quantization methods on Llama3-8b across different benchmarks.

Causal Mechanisms and Temporal Analysis of Concentrated Massive Values

绝对位置编码简要介绍

目前的位置编码主要分为绝对位置编码和相对位置编码两种

绝对位置编码：**为每个位置分配一个固定的位置编码，并与word embedding相加**

- 训练式位置编码：每个位置的位置向量会随模型一起训练，词向量直接与位置向量相加
  - 代表模型：BERT, GPT1/2/3
  - 缺点: 不具有长度外推性 $$q_m = f(q, m) = q + p_m$$
- Sinusoidal位置编码: pos表示位置索引，2i和2i+1表示位置向量的分量索引

$$PE_{(pos,2i)} = \sin\left(pos/10000^{2i/d}\right)$$
$$PE_{(pos,2i+1)} = \cos\left(pos/10000^{2i/d}\right)$$

Sinusoidal位置编码特殊性质

- **相对位置表达能力**: 对于固定的位置距离k, PE(t+k)可以表示成PE(t)的线性函数

$$
\begin{aligned}
PE\left(t,2i\right) &= sin\left(t * w_{2i}\right) \\
PE\left(t,2i+1\right) &= cos\left(t * w_{2i}\right) \\
w_{2i} &= 1/10000^{2i/d}
\end{aligned} \tag{2}
$$

$$
\begin{aligned}
PE\left(t+k,2i\right) &= sin\left(t * w_{2i} + k * w_{2i}\right) \\
&= sin\left(t * w_{2i}\right) cos\left(k * w_{2i}\right) + cos\left(t * w_{2i}\right) sin\left(k * w_{2i}\right) \\
&= PE\left(t,2i\right) PE\left(k,2i+1\right) + PE\left(t,2i+1\right) PE\left(k,2i\right) \\
&= PE\left(t,2i\right) u + PE\left(t,2i+1\right) v \\
PE\left(t+k,2i+1\right) &= cos\left(t * w_{2i} + k * w_{2i}\right) \\
&= cos\left(t * w_{2i}\right) cos\left(k * w_{2i}\right) - sin\left(t * w_{2i}\right) sin\left(k * w_{2i}\right) \\
&= PE\left(t,2i+1\right) PE\left(k,2i+1\right) - PE\left(t,2i\right) PE\left(k,2i\right) \\
&= PE\left(t,2i+1\right) u - PE\left(t,2i\right) v
\end{aligned} \tag{3,4}
$$

和t无关，是关于k的常数

仅使用位置t时的两维就可以得到位置t+k对应的两维

Sinusoidal位置编码特殊性质

- **相对位置表达能力**: 两个位置向量的**内积只和相对位置k有关**

$$
\begin{aligned}
PE(t)\,PE(t+k) &= \sum_{i=0}^{d/2-1} PE(t,2i)\,PE(t+k,2i) + \sum_{i=0}^{d/2-1} PE(t,2i+1)\,PE(t+k,2i+1) \\
&= \sum_{i=0}^{d/2-1} sin\left(t * w_{2i}\right) sin\left[(t+k) * w_{2i}\right] + \sum_{i=0}^{d/2-1} cos\left(t * w_{2i}\right) cos\left[(t+k) * w_{2i}\right] \\
&= \sum_{i=0}^{d/2-1} cos\left(k * w_{2i}\right)
\end{aligned}
$$

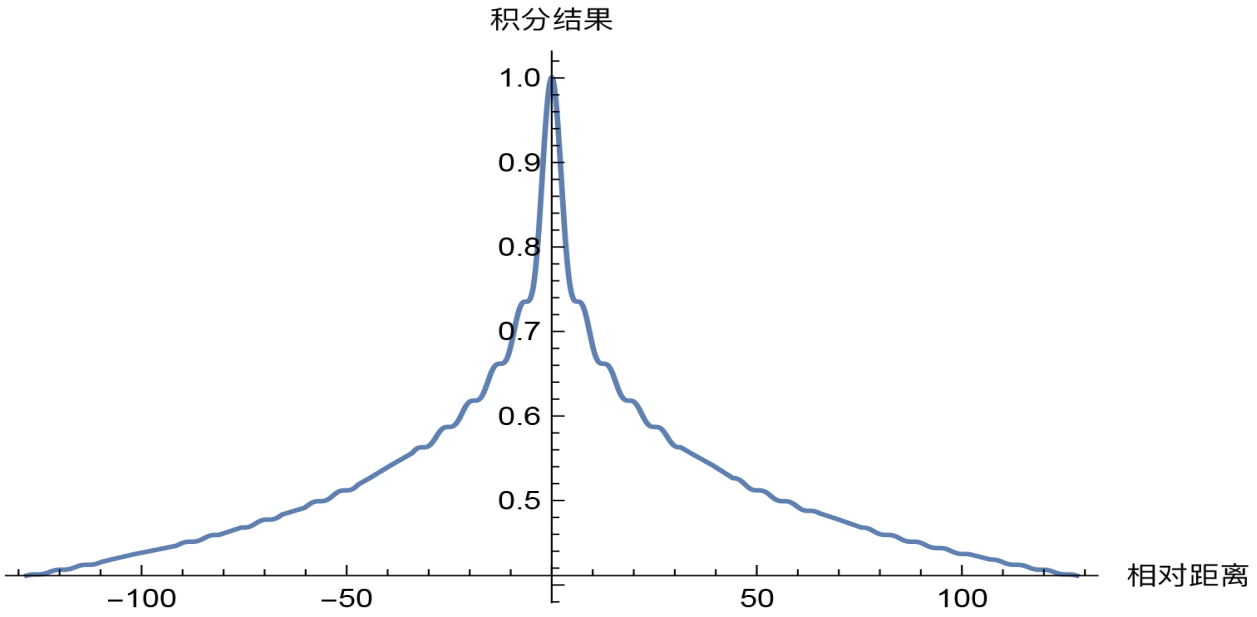
- **周期性**: **重要性质，ROPE的设计参考这个性质**
  - i小的分母小，周期小，为高频维度。**高频维度精确区分邻近位置**
  - i大的分母大，周期大，为低频维度。**低频维度表示全局信息**

$$
PE_{(pos,2i)} = \sin\left(pos/10000^{2i/d}\right)
$$

$$
PE_{(pos,2i+1)} = \cos\left(pos/10000^{2i/d}\right)
$$

Sinusoidal位置编码特殊性质

- **远程衰减**：随相对位置的增加，位置向量的内积会逐渐降低
  - 离得远的注意力权重应该小



- 对称性: PE(t+k)PE(t) = PE(t)PE(t−k)
- 相比于可学习的绝对位置编码，Sinusoidal位置编码具有**外推性**

Sinusoidal位置编码-理论与实际的差距

- 理论: 上述的所有性质都是在仅考虑位置编码的情况下得到的

- 实际：很多信息揉在一起，学习缺乏显式的指导
  - 位置编码需要与word embedding相加。虽然理论上点积能捕捉相对位置，但这依赖于模型自己从混合信号中学习，难
  - embedding还需要与转换矩阵相乘，变成Q和K

建模相对位置信息对外推性很重要，怎么做？

相对位置编码简要介绍

相对位置编码：考虑相对位置关系，通常不直接把位置编码加到word embedding上

- **修改注意力分数计算**：在Q和K计算完分数之后，显式地加上一个只依赖于相对距离 k = n - m 的偏置项 b_{k}
  - AttentionScore_{m, n} = (Q_m · K_n) + b_{n-m}
  - 对于不同范围的k，可以设置不同的b_k，b_k可以是可学习的参数
  - 代表模型: Transformer-XL, T5, ALiBi

- **Rotary Position Embedding (RoPE)**：通过旋转Q和K，以绝对位置编码的形式实现相对位置编码
  - 代表模型: Llama, GLM, Baichuan, Qwen等

RoPE

主要思想：通过直接旋转Q和K来实现相对位置编码

- 希望q_m和k_n之间的点积能够带有相对位置信息 (m-n)

$$f(q,m)\cdot f(k,n)=g(q,k,m-n)$$

对位置m和n的位置编码

以二维为例，最终得到位置编码如下：R_m是一个旋转矩阵，逆时针旋转m\theta度

- 通过旋转为一个向量添加绝对位置信息，这就是旋转位置编码的由来

$$f(q,m)=R_m q=\begin{pmatrix}\cos m\theta & -\sin m\theta\\ \sin m\theta & \cos m\theta\end{pmatrix}\begin{pmatrix}q_0\\ q_1\end{pmatrix}$$

RoPE

主要思想：通过直接旋转Q和K来实现相对位置编码

- q_m和k_n之间的点积能不能捕捉相关位置信息？ **能!**

$$
\begin{aligned}
q_m \cdot k_n &= f(q,m) \cdot f(k,n) = (R_m q)^T * (R_n k) = q^T R_m^T * R_n k \\
&= q^T \begin{bmatrix} \cos m\theta & -\sin m\theta \\ \sin m\theta & \cos m\theta \end{bmatrix}^T * \begin{bmatrix} \cos n\theta & -\sin n\theta \\ \sin n\theta & \cos n\theta \end{bmatrix} k \\
&= q^T \begin{bmatrix} \cos m\theta & \sin m\theta \\ -\sin m\theta & \cos m\theta \end{bmatrix} * \begin{bmatrix} \cos n\theta & -\sin n\theta \\ \sin n\theta & \cos n\theta \end{bmatrix} k \\
&= q^T \begin{bmatrix} \cos n\theta \cos m\theta + \sin n\theta \sin m\theta & \sin m\theta \cos n\theta - \sin n\theta \cos m\theta \\ \sin n\theta \cos m\theta - \sin m\theta \cos n\theta & \cos n\theta \cos m\theta + \sin n\theta \sin m\theta \end{bmatrix} k \\
&= q^T \begin{bmatrix} \cos(n-m)\theta & -\sin(n-m)\theta \\ \sin(n-m)\theta & \cos(n-m)\theta \end{bmatrix} k \\
&= q^T R_{n-m} k
\end{aligned}
$$

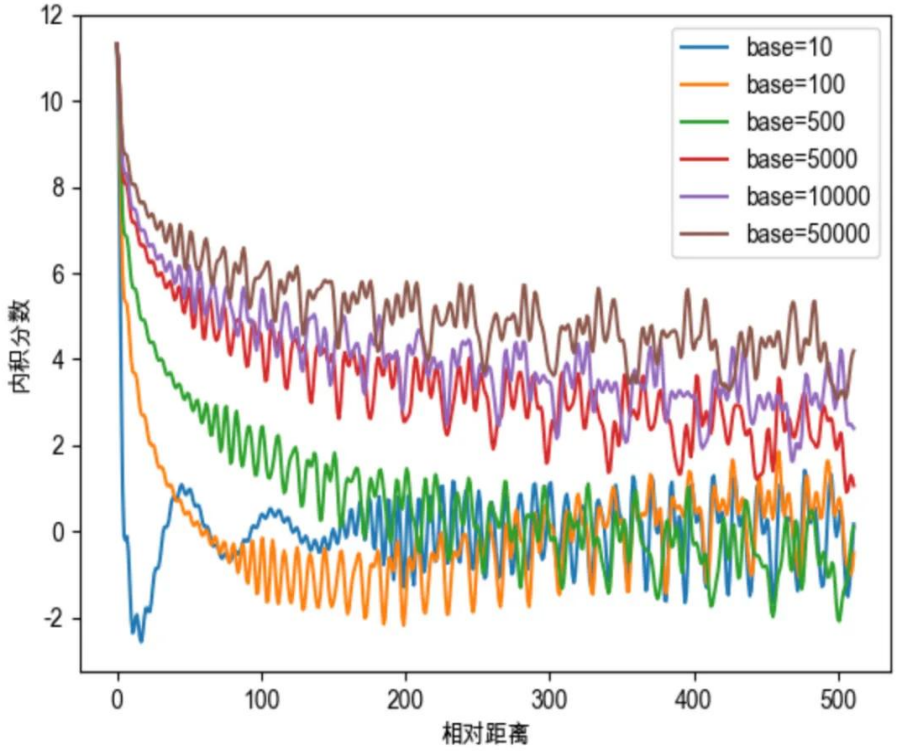
RoPE

从二维推广到高维：**两两一组，分别旋转，每一组有不同的旋转周期**

- **周期小，转速快的区分邻近位置**
- **周期大，转速慢的区分整体位置**

$$\boldsymbol{R}^d_{\Theta,m} = \underbrace{\begin{pmatrix} \cos m\theta_0 & -\sin m\theta_0 & 0 & 0 & \cdots & 0 & 0 \\ \sin m\theta_0 & \cos m\theta_0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cos m\theta_1 & -\sin m\theta_1 & \cdots & 0 & 0 \\ 0 & 0 & \sin m\theta_1 & \cos m\theta_1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & \cos m\theta_{d/2-1} & -\sin m\theta_{d/2-1} \\ 0 & 0 & 0 & 0 & \cdots & \sin m\theta_{d/2-1} & \cos m\theta_{d/2-1} \end{pmatrix}}_{\boldsymbol{W}_m}$$

$$\Theta = \left\{\theta_i = 10000^{-2i/d}, i \in [0, 1, \ldots, d/2-1]\right\}$$



- 有远程衰减性

RoPE Contributes to Concentrated Massive Values

- Evidence1: This concentration of massive values in low-frequency regions primarily encodes rich semantic content rather than positional information, as disrupting these values severely impairs contextual understanding tasks

Model	RoPE	Concentrated Massive Values
Llama 2, 3	✓	✓
Qwen 2, 2.5	✓	✓
Gemma 1, 2	✓	✓
Phi-3	✓	✓
Falcon3	✓	✓
LLAVA	✓	✓
Qwen2-VL	✓	✓
Mistral-v0.3	✓	✓
GPT-NeoX	✓	✓
GPT-2	×	×
GPT-Neo	×	×
OPT all size	×	×
Jamba	×	×

Table 3. Analysis of different models about whether they adopt RoPE and whether concentrated massive values in Q and K can be observed in these models.

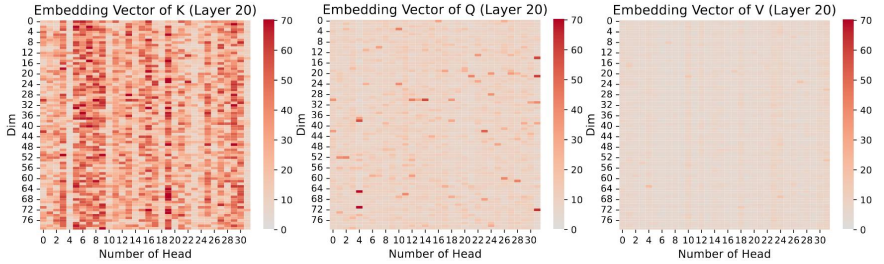


Figure 6. K, Q and V in Layer 20 of OPT-350M, where massive values are not concentrated in specific regions.

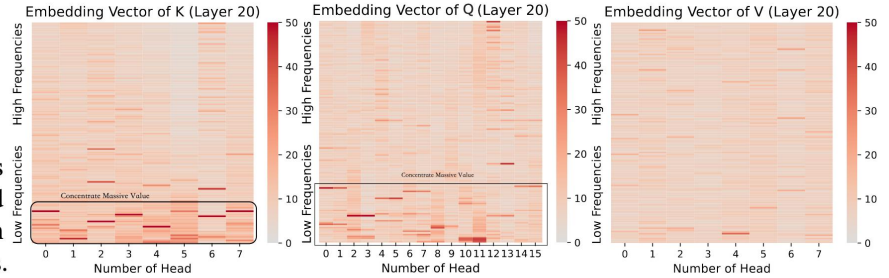


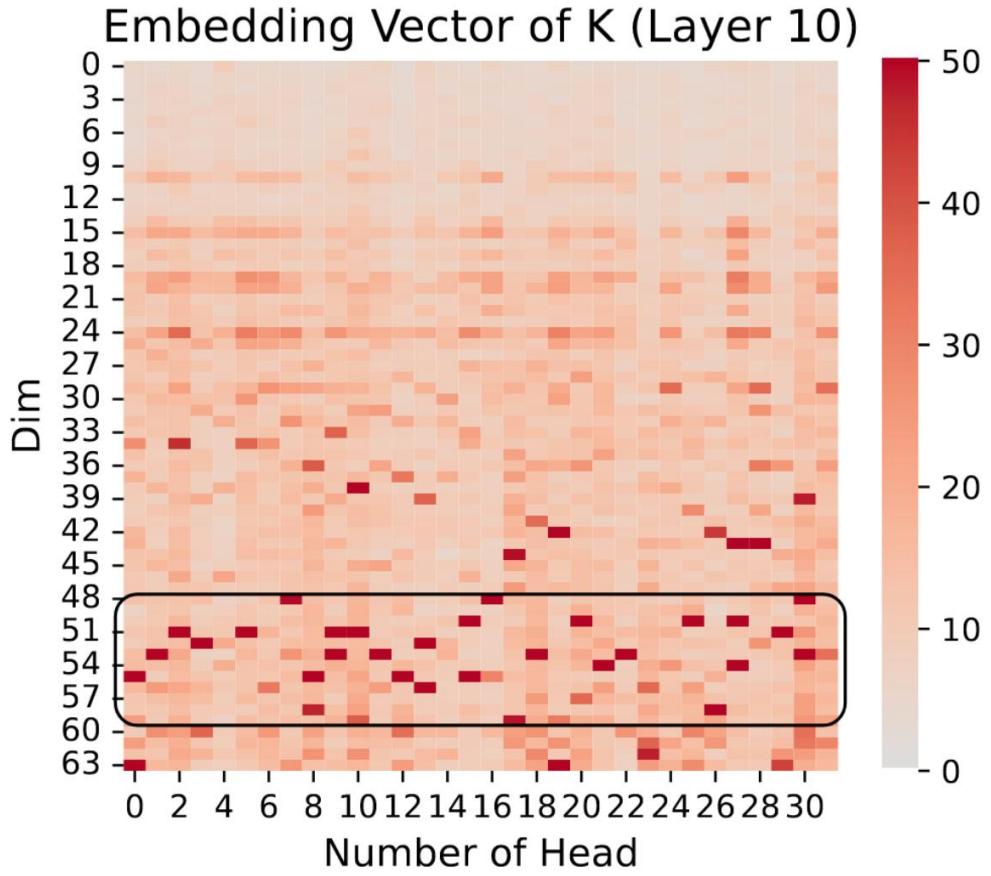
Figure 7. K, Q, and V at Layer 20 in Gemma2-9B. Massive values are concentrated in low-frequency regions of K and Q, absent in V.

- Evidence2: RoPE applies position encoding on K and Q but not on V. The **concentrated** massive value appear exclusively in the Q and K, while being completely absent in V.

- **Evidence3**: Concentrated Massive Value in Q and K appears exclusively in the LLM with RoPE like Gemma

RoPE Contributes to Concentrated Massive Values

- Evidence4: Observe two distinct clusters of massive values in the embedding vector, with one cluster appearing in the first half of the dimensions and a corresponding cluster in the second half, creating **a symmetric pattern**.



我没看出来哪对称，文中也没说明哪个图能看出来对称

RoPE Contributes to Concentrated Massive Values

- LLMs exhibit massive values in Q and K starting from the very first layer.

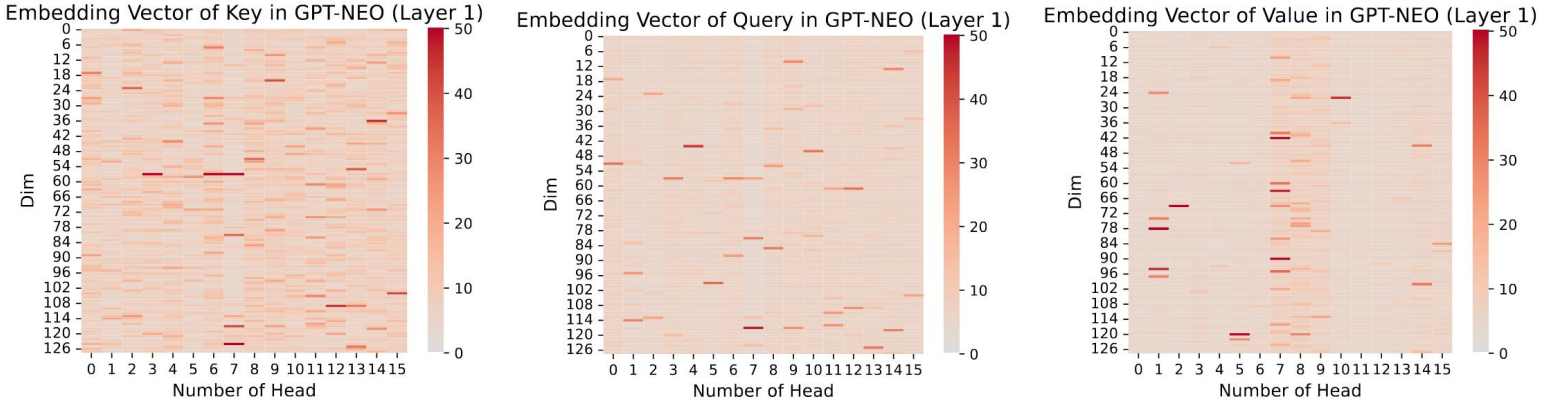


Figure 44. Embedding Vector of K Q, V in GPT-NEO-1.3B without RoPE, we choose Layer 1 and the input question is shown as Figure 11

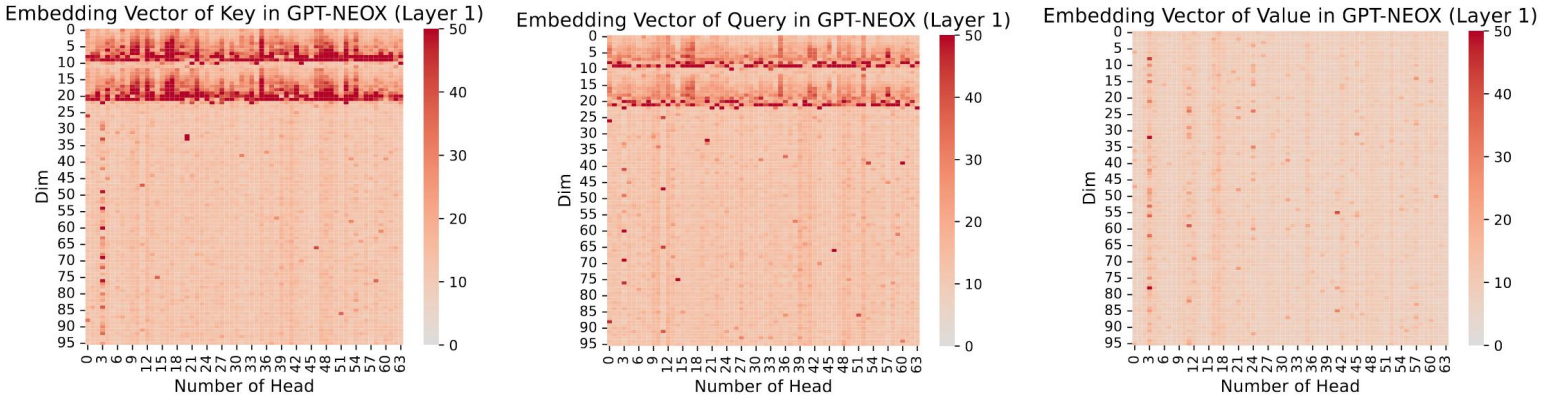


Figure 45. Embedding Vector of K Q, V in GPT-NEOX-20B with RoPE, but not the same RoPE as Llama/Gemma), we choose Layer 1 and the input question is shown as Figure 11

Takeaway

- Massive values are concentrated in **specific regions** of Q and K exclusively.
  - These massive values in each head's dim index are very close
  - Absent in V and absent in models without RoPE

- Massive values in Q and K are **critical for understanding contextual knowledge** over parametric knowledge
  - Disrupting these values leads to a notable degradation in tasks requiring contextual understanding

- Quantization techniques **targeting massive values** preserve contextual knowledge better

- Concentration of massive values is **caused by RoPE** and it **appears since very early layers** in LLMs

Thanks & QA

参考链接

- https://spaces.ac.cn/archives/8231
- https://www.zhihu.com/tardis/zm/art/675243992?source_id=1003
- https://mp.weixin.qq.com/s/-1xVXjoM0imXMC7DKqo-Gw