Found in the Middle: How Language Models Use Long Contexts Better via Plug-and-Play Positional Encoding

Zhenyu Zhang[1*], **Runjin Chen**[1], **Shiwei Liu**[2], **Zhewei Yao**[3], **Olatunji Ruwase**[3], **Beidi Chen**[4], **Xiaoxia Wu**[3†], **Zhangyang Wang**[1†]
[1]University of Texas at Austin, [2]University of Oxford, [3]Microsoft, [4]Carnegie Mellon University
[*] Work done during internship at Microsoft, [†] Equal advising

Abstract

This paper aims to overcome the "lost-in-the-middle" challenge of large language models (LLMs). While recent advancements have successfully enabled LLMs to perform stable language modeling with up to 4 million tokens, the persistent difficulty faced by most LLMs in identifying relevant information situated in the middle of the context has not been adequately tackled. To address this problem, this paper introduces Multi-scale Positional Encoding (Ms-PoE) which is a simple yet effective plug-and-play approach to enhance the capacity of LLMs to handle the relevant information located in the middle of the context, without fine-tuning or introducing any additional overhead. Ms-PoE leverages the position indice rescaling to relieve the long-term decay effect introduced by RoPE, while meticulously assigning distinct scaling ratios to different attention heads to preserve essential knowledge learned during the pre-training step, forming a multi-scale context fusion from short to long distance. Extensive experiments with a wide range of LLMs demonstrate the efficacy of our approach. Notably, Ms-PoE achieves an average accuracy gain of up to 3.8 on the Zero-SCROLLS benchmark over the original LLMs. Code are available at `https://github.com/VITA-Group/Ms-PoE`.

1 Introduction

Effective long-sequence reasoning in large language models (LLMs) is crucial for a wide range of applications [1, 2], from understanding extensive texts [3, 4] and managing day-long conversations [5, 6] to code generation [7, 8] and science discoveries [9, 10]. Recent system support advancements [11, 12] have enabled training transformers for any L sequence length even with $O(L^2)$ computational complexity. This is exemplified by models such as MPT [13] and Mistral [14] pre-trained with sequence lengths 16k and 32k respectively.

Nevertheless, emerging research reveals the constrained efficacy of LLMs in managing tasks requiring long contextual understanding. Particularly, [15] demonstrated a substantial degradation in LLMs' performance when crucial information is positioned amidst a lengthy context, a phenomenon they refer to as "lost-in-the-middle". One explanation is about the use of rotary positional embedding (RoPE) [16], a prevalent positional encoding technique used in open-source LLMs. As a relative position embedding, RoPE incorporates a long-term decay property, predisposing the model to prioritize current/nearby tokens while paying less attention to further ones. [17] identified a surprising trend attributed to the Softmax operation where attention scores are disproportionately allocated into initial tokens, irrespective of their relevance to the language modeling task. Despite the presence of considerable redundancy in long-context inputs [18], crucial information may be located across different positions. The inclination of LLMs to overlook the middle section presents a challenge for their applications, particularly in the context of long-context reasoning. Several approaches successfully extend pre-trained LLMs with context up to extreme token length, either through sparse

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

 https://doi.org/10.52202/079017-1943

selection of crucial tokens during generation [17, 18, 19] or by modifying positional encoding [20, 21]. *Nevertheless, these approaches primarily aim to extend the context length of LLMs and, consequently, fall short in addressing the "lost-in-the-middle" problem when applied out-of-the-box.*

Efforts have been made to enhance LLMs' capacity to capture vital information located within the middle of the context. These include extra memory bank [22], reordering the input context based on relevance [23, 24], enhancing the information searching and reflection ability via attention strengthening tasks [25, 26], splitting the input into short segments and applying short-text models [27]. For example, [23] empirically discovered that LLMs tend to emphasize more on the current window while still paying more attention to the relevant text than distracting content. They subsequently introduced "attention sorting" where the main idea is iteratively sorting documents based on their attention scores, such that critical information will likely be placed at the end, to fit the position-biased nature of RoPE. [24] conducted parallel runs of LLMs with different RoPE angles, thereby mitigating the risk of overlooking crucial information through a weighted sum of the outputs. These approaches usually require additional memory or multiple inference runs, which can be expensive for LLMs.

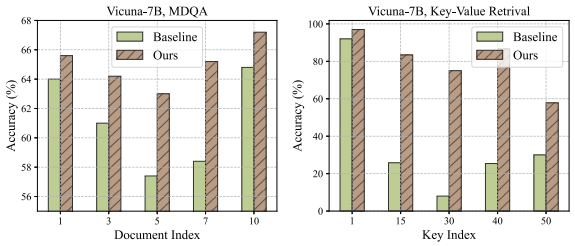

Figure 1: The x-axis illustrates the placement of essential information within the prompt, ranging from start to end. The green bar serves as a standard baseline, illustrating the "lost-in-the-middle" phenomenon. We introduce our method, **M**ulti-**s**cale **Po**sition **E**ncoding (Ms-PoE), which requires neither additional fine-tuning nor increased memory usage. Instead, it involves a simple remapping of the position embedding depicted in Figure 2, which enables the important information in the middle to be detected effectively (brown bars). For more details, see Section 4.2 and Figure 5.

In this paper, we aim to address the "lost-in-the-middle" problem by reintroducing the concept of multi-scale features from computer vision into the context of Transformer-based LLMs. Multi-scale features, well-established in Inception-style models [28, 29, 30], utilize parallel employment of kernels with different sizes to fuse multi-scale information, spanning short to long distances. Introducing multi-scale operations into LLMs intuitively can help compensate for crucial information located in the middle, which might be easily overlooked by full attention operation. Unlike modifying the attention module to form multi-scale attention, we choose to re-scale the indices of positional encoding. This decision is grounded not only in its effectiveness in easily adjusting the scale of the context window by simply changing the position indices [20] but also in the potential of down-scaling the position indices to relieve the long-term decay property introduced by RoPE. However, this approach was initially introduced to extend context windows, and its performance regarding the "lost-in-the-middle" problem remains uncertain for several reasons: (i) Indice re-scaling forces position embeddings of original context window to reside in a narrower region, leading to performance degradation in the original context window as shown in [20]. (ii) Uniformly applying the same scaling ratio throughout the entire model might be sub-optimal to preserve essential knowledge learned during pre-training; (ii) Fine-tuning is necessary for the original approach, albeit minimal. The impact without fine-tuning remains unknown.

To this end, we systematically visit the position indices scaling regarding the "lost-in-the-middle" problem and counter-intuitively discover that it is possible to slightly mitigate the "lost-in-the-middle" issue if we carefully choose the scaling ratio to be around 1.5-2. Additionally, we observe that different attention heads exhibit varying sensitivity to the position shift of the relevant document. Some attention heads are "position-aware", consistently capturing relevant information even with position shifts, while others may occasionally capture position changes, and some heads are completely insensitive to position changes. This highlights the need to treat attention heads separately when re-scaling position indices.

Contribution. Inspired by the above observations, we introduce Multi-scale Positional Encoding (Ms-PoE), a simple yet effective plug-and-play approach that can enhance the long-context reasoning capability of pre-trained LLMs without requiring fine-tuning or introducing any additional overhead. Ms-PoE meticulously assigns distinct scaling ratios to different attention heads, with the scaling factor monotonically increasing from "position-aware" heads to "position-unaware" heads. This enables

https://doi.org/10.52202/079017-1943

us to improve long-context ability by re-scaling position indices to shorter values while preserving essential knowledge acquired during the pre-training phase. The efficacy of Ms-PoE is substantiated through extensive experiments. By simply re-scaling the indices of positional encoding, Ms-PoE consistently enhances the performance of various LLMs including Llama-2 [31], StableBeluga [32] and Vicuna [33] on the ZeroSCROLLS [34], achieving a notable accuracy gain of up to 3.8.

2 Background and Related Works

In this section, we provide a concise overview of the background knowledge and recent literature about the generative inference process of Large Language Models (LLMs), their abilities for long-context reasoning, and details of positional encoding.

2.1 Generative Inference of LLMs

The generative inference process in LLMs can be categorized into two distinct phases: ① Prefilling Stage: In this initial phase, LLMs receive an input sequence containing detailed instructions that define a specific generation goal. Throughout this stage, intermediate Key and Value embeddings are generated at each layer and stored in memory, commonly referred to as the KV cache. ② Decoding Stage: This phase involves retrieving embeddings from the KV cache to generate new tokens. The decoding process is inherently iterative, where each newly generated token serves as input for the subsequent token generation. In real-world LLM deployment, the cumulative length of input sequences and the subsequently generated text can reach several thousand or even millions of tokens, presenting significant challenges for the LLMs' long-context reasoning capability.

2.2 Long Context Reasoning

Two challenges for LLMs in handling long-context reasoning tasks. One is to extend the context window to process sentences that exceed the pre-trained window length. Another is the "lost-in-the-window" issue where LLMs likely overlook the information located in the middle of the sentences.

The reason for the former challenge is that open-source LLMs are usually pre-trained with fixed sequence lengths, such as 4096 for Llama-2 [31]. When the sequence length surpasses the predefined context length used in pre-training, LLMs often suffer from performance collapses and thus generate incoherent or fragmented text. Recent efforts to address this issue can be broadly categorized into two streams. Recently, several works have been proposed to address this issue, which can be broadly categorized into two streams. The first one explores from the expansion of positional encoding, with notable contributions including PI [20], CLEX [35], YaRN [36], Self-Extend [21]. On the other hand, some works modify the attention mechanism, such as StreamingLLM [17], LM-Inifinite [19], H_2O [18], TOVA [37], Zebra [38], and Activation Beacon [39]. These approaches have successfully expanded the contextual window with minimal or no additional training overhead.

Despite the extended context window, LLMs still face a significant challenge in long-context inference due to the uneven utilization of lengthy inputs. [15] conducted a pivotal investigation, revealing that LLMs tend to overlook the middle portion of the input. This bias compromises the practical application of LLMs, as critical information may be located in the middle part of the input, leading to unreliable outputs. To tackle this issue, [23] introduced 'attention sorting' to reorder inputs, placing critical information at the end. However, this method's reliance on potentially biased attention scores to identify crucial content may compromise its reliability, and the prerequisite knowledge of document count in inputs may affect its effectiveness. [24] utilize Attention Buckets, an ensemble approach that combines multiple forward processes with positional modifications. However, this technique necessitates a considerably higher computational cost. Other general approaches for enhancing long-context reasoning include prompt compression [40], retrieval augmentation [26], and inference refinement by constructing memory trees [41] while these approaches typically necessitate extra LLMs' assistance or bring extra computational cost.

2.3 Positional Encoding

For effective processing of long contexts, LLMs necessitate the explicit encoding of positional information. Common techniques include absolute positional embedding and relative positional encoding. Absolute positional embedding integrates word embeddings with an additional positional

 https://doi.org/10.52202/079017-1943

vector based on the token's absolute position, which can be either fixed [42] or learnable [43, 44, 45, 46, 47]. In contrast, relative positional encoding, increasingly popular in contemporary LLMs, encodes the relative distances between tokens instead of their absolute positions. Notable among these are Rotary Position Embedding (RoPE) [16] that widely implemented in models like Llama [31], Falcon [48], Mistral [49], and ALiBi [50], which used in MPT [13].

RoPE. The primary goal of RoPE [16] is to encode positional information such that the inner product of the query and key embeddings inherently contains the relative position information:

$$f(\mathbf{q}_m, m)^T f(\mathbf{k}_n, n) = g(\mathbf{q}_m, \mathbf{k}_n, m-n)$$

Here, f is the positional encoding function applied to the query and key embeddings at positions m and n, respectively. To satisfy this condition, the function f is defined as a vector-valued complex function, as follows:

$$\begin{aligned} f(\mathbf{x}, m) &= \mathbf{x}e^{im\theta} \\ &= [(x_1 + ix_2)e^{im\theta_1}, (x_3 + ix_4)e^{im\theta_2}, \\ &\ldots, (x_{l-1} + ix_l)e^{im\theta_{l/2}}]^T \end{aligned}$$

In this equation, l represents the dimension of the embeddings, $\theta_k = 10000^{-2k/l}$, and i is the imaginary unit. For calculating the attention score, RoPE considers the real part of the product, specifically $\mathrm{Re}(f(\mathbf{q}_m, m)^T f(\mathbf{k}_n, n))$. This approach allows RoPE to effectively integrate relative positional information into the attention mechanism of transformer models.

3 Methodology

In this section, we present the details of our Multi-Scale Positional Encoding (Ms-PoE) approach. Section 3.1 demonstrates that the context utilization of LLMs can be directly enhanced by re-scaling the positional information without incurring extra training costs. Then, Section 3.2 analyzes the properties of various attention heads in LLMs. Section 3.3 outlines the detailed pipeline of Ms-PoE.

3.1 Positional Re-scaling Improves Context Utilization

Current LLMs tend to neglect information located in the middle of the context, despite its potential relevance. This "lost in the middle" phenomenon likely arises from two contributing factors: (i) Casual Attention, where preceding tokens undergo a higher number of attention processes, leading LLMs to disproportionately favor initial tokens. This phenomenon has been demonstrated in recent research which highlights the pivotal role of the initial tokens in model generation [19, 17], with these starting tokens consistently accumulating higher attention scores [18]. (ii) The utilization of RoPE [16] introduces a long-term decay effect, diminishing the attention score of distantly positioned yet semantically meaningful tokens. The combination of these factors contributes to LLMs neglecting the context in the middle part. To tackle this issue and improve the context utilization of LLMs, a seemingly unreasonable yet remarkably effective strategy is to down-scale positional information [38]. Formally, RoPE encodes the position as $f(\mathbf{x}, m) = \mathbf{x}e^{im\theta}$. By substituting the position m with $\frac{m}{r}$, we can force the long-distance tokens to

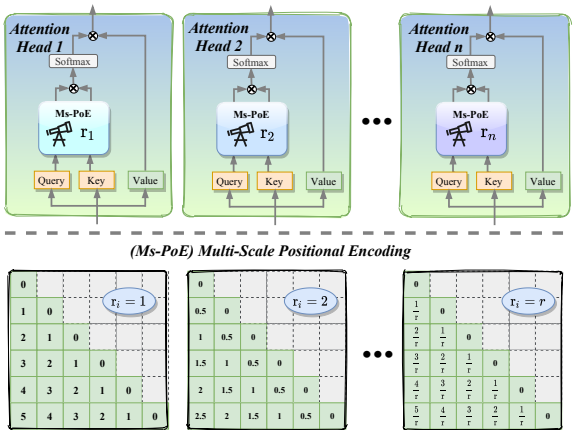

Figure 2: Illustration of our Multi-scale Positional Encoding (Ms-PoE) framework. The top figure demonstrates the implementation of Ms-PoE with various scaling ratios in different attention heads, marked with different colors. The bottom figure shows the position details of each head, in which the first matrix ($r_i = 1$) represents the original RoPE.

reside in the shorted range, which can potentially alleviate the long-term decay effects by a factor of r. In the following sections, we conduct experiments to evaluate how LLMs' context utilization responds to varying re-scaling ratios r.

Details. Experiments are conducted using Llama-2-7B-Chat [31] and Vicuna-7B [33] on the Multi-Document Question Answering (MDQA) task [15]. Each question includes ten documents, with only one relevant to the question. By varying the position of the relevant document, we can evaluate LLMs' context utilization properties. For each position of the key document, we calculate the accuracy over 500 samples. And results show in Figure 3 include both the **Average** accuracy over the 10 documents as well as **Gap** accuracy, *i.e.*, the difference between the best and worst accuracy when varying the positions of the relevant document.

Results. Figure 3 demonstrates that the gap accuracy can be alleviated via appropriate positional re-scaling. Particularly, we see that the Gap between the best and the worst accuracy is greatly reduced when increasing the re-scaling ratio. An enhanced average accuracy can be observed with a scaling ratio equals near 1.5. Additionally, changing the scaling ratio also affects the favored zone of LLMs. With a small scaling ratio (e.g., 0.5), LLMs tend to focus more on the most recent part of the context, while with a large ratio (e.g., 2.5), LLMs favour the beginning part.

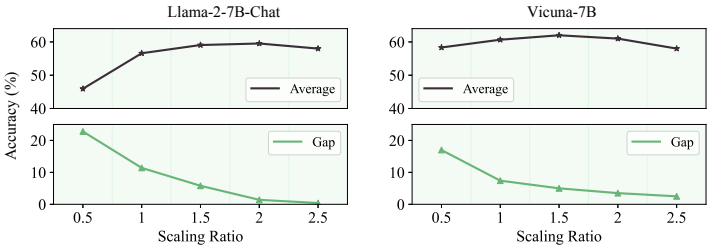

Figure 3: Results of the relationship between positional re-scaling and context utilization. The upper curve illustrates the average accuracy when placing the key document in various positions. The bottom curve indicates the gap between the best and worst accuracy.

Improving context reasoning via positional re-scaling. Building upon this, we introduce a plug-and-play treatment for RoPE by re-scaling the position of each token. This approach seamlessly enhances the context utilization of LLMs without requiring additional training or inference overhead. However, there is a trade-off in terms of LLMs favoring certain context regions. For instance, when $r = 0.5$, LLMs achieve peak accuracy when the relevant document is located at the end of the input, while at the beginning for $r = 1.5$. It remains challenging to decide which re-scaling ratio to use, given that we lack prior knowledge of the location of relevant information in real-world applications. Moreover, as the re-scaling ratio increases, LLMs may face the positional out-of-distribution (O.O.D) issue [21, 20], where many position values do not directly exist during pretraining (e.g., using $0.1, 0.2, ..., 0.9$ for position when LLMs only recognize $1, 2, ..., 9$ during pretraining), potentially reducing their average reasoning ability. To tackle these challenges, we investigate the head-wise properties of LLMs and propose a multi-scale positional encoding approach.

3.2 Position-Aware Head-Wise Re-scaling Ratio

Inspired by recent works that leverage attention patterns to identify most crucial tokens and optimize inference efficiency [37, 18, 51], we carry out a preliminary study to investigate the interaction between attention patterns and token positions.

Details. We visualize the attention patterns of the most recent query with results collected from Vicuna-7B on the MDQA task, following [37]. In the same input sample, we manually switch the position of the relevant document from the beginning to the end and illustrate the attention scores across different positions.

Observation. We observe the presence of "position-aware" attention heads capable of capturing relevant information even when its position is shifted. As an example, we select the eighth attention head in the fifteenth layer, depicted in the bottom of Figure 4, while consistent observations can be drawn across different layers and input samples. Firstly, most attention scores are near zero and can be ignored, consistent with other studies highlighting high sparsity in attention blocks [18, 52, 53]. For the remaining positions, these "position-aware" attention heads can capture important information across positions, with attention patterns shifting as the position of relevant tokens changes. However,

 https://doi.org/10.52202/079017-1943

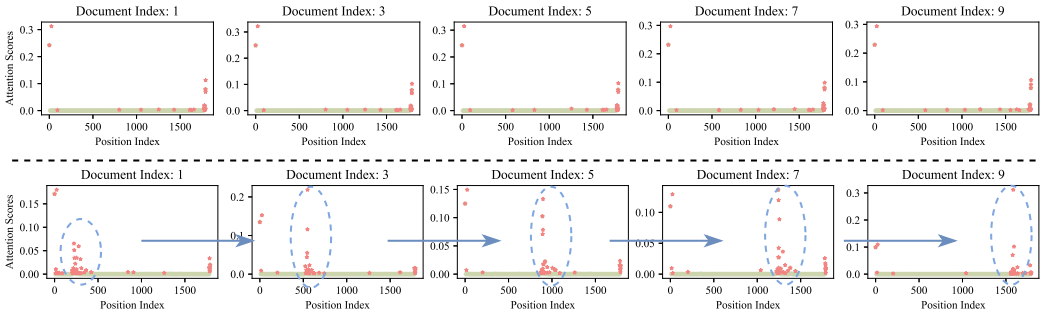

Figure 4: Visualization of attention pattern of the most recent query within two different attention heads. **Top:** Results of the 12th attention head in the 15th layer. **Bottom:** Results of the 8th attention head in the 15th layer. The most recent query remains unchanged while varying the position of the crucial document. More examples are reported in Figure 6 in the appendix.

for other attention heads (upper subfigure in Figure 4), they fail to capture relevant tokens and only attend to the beginning and end words, contributing to the "lost-in-the-middle" issue.

Based on this observation, we devise a position-aware strategy to adaptively determine the re-scaling ratio via the inherent properties of different attention heads. For the "position-aware" attention heads, we assign a re-scaling ratio close to one to avoid changing their functionality significantly, as altering them too much could degrade performance due to the positional O.O.D issue. On the other heads, we condense their position indices to a higher degree, providing more opportunity to alleviate the persistent bias toward the beginning and recent tokens. To identify the properties of n_h attention heads, we introduce a Position-Awareness Score $\mathcal{S}_P \in \mathbf{R}^{n_h}$ formulated as:

$$\mathcal{S}_P = \frac{1}{l}\sum_{i=1}^{l}(A_i \geq \alpha \frac{1}{l}\sum_{i=1}^{l} A_i) \tag{1}$$

In Equation 1, A represents the attention score vector of the most recent query, and α is a hyper-parameter determining the threshold of effective attention scores. In all experiments, we default to using $\alpha = 3$, and the corresponding important tokens are highlighted in Figure 4, which are shown in red. In the spirit of numerous studies that investigate the outlier properties in LLMs [17, 54, 55], we utilize $\mathcal{S}_P$ to evaluate the ratio of effective attention tokens, where a larger $\mathcal{S}_P$ value implies better positional awareness.

3.3 Inference with Multi-Scale Positional Encoding

The pipeline for utilizing Multi-Scale Positional Encoding (Ms-PoE) in LLM inference is: Given a pre-trained LLM, we initially replace the original rotary positional encoding with Ms-PoE. As illustrated in Figure 2, Ms-PoE condenses the positional indices of RoPE and employs different re-scaling ratios for each attention head. The re-scaling ratios are assigned during the prefilling stage, where we first calculate the distribution of attention scores for the most recent query and obtain the corresponding position-awareness score for each attention head. Larger re-scaling ratios are subsequently allocated to attention heads exhibiting smaller position-awareness scores. And the set of re-scaling ratios $\mathbf{r}$ defaults to a **linear** range from R_{min} to R_{max}. For example, the ith sorted-head would be using re-scaling ratio

$$r_i = R_{min} + (i-1)(R_{max} - R_{min})/(n_h - 1) \tag{2}$$

Once the re-scaling ratios are assigned, they remain fixed in the subsequent decoding stage. We consistently using $R_{min} = 1.2$ and $R_{max} = 1.8$ is our experiments.

4 Experiments

The goal of this section is to demonstrate Ms-PoE, a plug-and-play positional encoding capable of enhancing the context utilization of LLMs, and consequently improving the quality of generation across diverse models and downstream reasoning tasks. Our main results can be summarized below.

Table 1: Comparsion results on ZeroSCROLLS [34] benchmarks. The evaluation metrics for various tasks are tailored as follows: GovReport, SummScreenFD, QMSum, and SQuALITY utilize the geometric mean of Rouge-1/2/L scores. Qasper and NarrativeQA are assessed through the F1 score, while BookSumSort employs the concordance index.

Models	Methods	GovReport	SummScreenFD	QMSum	SQuALITY	Qasper	NarrativeQA	BookSumSort	Average
Llama-2-7B-Chat	Baseline	16.8	14.1	15.2	19.5	21.9	14.4	3.1	15.0
Llama-2-7B-Chat	Ours	**17.7** (+0.9)	**14.2** (+0.1)	**15.8** (+0.6)	**19.9** (+0.4)	**25.1** (+3.2)	**17.7** (+3.3)	**5.8** (+2.7)	**16.6** (+1.6)
Llama-2-13B-Chat	Baseline	15.4	12.3	15.1	18.9	19.0	15.0	5.7	14.5
Llama-2-13B-Chat	Ours	**16.5** (+1.1)	**13.1** (+0.8)	**15.5** (+0.4)	**19.2** (+0.3)	**20.8** (+1.8)	**17.0** (+2.0)	**5.9** (+0.2)	**15.4** (+0.9)
StableBeluga-7B	Baseline	14.9	13.8	14.7	17.9	28.1	16.8	9.2	16.5
StableBeluga-7B	Ours	**16.6** (+1.7)	**14.2** (+0.4)	**15.2** (+0.5)	**18.7** (+0.8)	**36.9** (+8.8)	**18.0** (+1.2)	**14.2** (+5.0)	**19.1** (+2.6)
StableBeluga-13B	Baseline	5.7	7.1	12.9	13.3	19.2	13.4	4.8	10.9
StableBeluga-13B	Ours	**7.4** (+1.7)	**7.4** (+0.3)	12.8 (-0.1)	13.2 (-0.1)	**20.8** (+1.6)	13.4 (+0)	**5.6** (+0.8)	**11.5** (+0.6)
Vicuna-7B	Baseline	16.2	13.7	15.1	18.9	24.3	13.7	3.3	15.0
Vicuna-7B	Ours	**20.2** (+4.0)	**14.5** (+1.8)	**15.4** (+0.3)	**19.8** (+0.9)	**34.7** (+13.4)	**16.2** (+2.5)	**10.5** (+7.2)	**18.8** (+3.8)
Vicuna-7B-16K	Baseline	20.2	13.9	16.2	20.1	32.3	18.8	29.9	21.6
Vicuna-7B-16K	Ours	**21.4** (+1.2)	**14.3** (+0.4)	16.2 (+0)	**20.2** (+0.1)	**37.8** (+5.5)	**21.0** (+2.2)	**43.3** (+13.4)	**24.9** (+3.3)

In Section 4.1, we demonstrate that Ms-PoE consistently enhances reasoning over long contexts for a range of tasks in the ZeroSCROLLS benchmarks [34], all without the need for additional training. Additionally, Ms-PoE exhibits superior performance when compared to other methods in the field, including PI [20] and Self-Extend [21]. Detailed comparison results are shown in Tables 1 and 2.

In section 4.2, we highlight that Ms-PoE improves the context utilization and achieves consistent improvement when varying the position of critical information, as shown in Figure 1 & 5.

In Section 4.3, we conduct multiple ablation studies to assess the effectiveness of Ms-PoE under different scaling ratios and selection strategies. Results are reported in Table 3 & 4.

4.1 Enhanced Generation Quality

We empirically validate the ability of Ms-PoE to enhance long-context reasoning with a noteworthy improvement up to 13.4 without additional training overhead. Notably, our approach surpasses other competitive baselines, demonstrating improvements from 2.64 to 43.72.

Experimental Setup. In our experiments, we select seven representative LLMs, including Llama-2-chat-7B and 13B [31], StableBeluga-7B and 13B [32], and Vicuna-7B [33], along with its longer-context version (Vicuna-7B-16K). To comprehensively evaluate the long-context reasoning abilities of LLMs, we choose seven tasks from ZeroSCROLLS [34], spanning all four task categories: ① Document Summarization (Government and SummScreenFD), ② Query-Based Summarization (QMSum and SQuALITY), ③ Question Answering (Qasper and NarrativeQA), and ④ Information Aggregation (BookSumSort). We also compare Ms-PoE with other competitive methods on additional generation tasks, including Multi-document Question Answering (MDQA) and Key-Value Retrieval [15].

Main Results. Table 1 summarizes the main results, yielding several key observations: (i) By simply substituting the original positional encoding module with our Ms-PoE, the performance of LLMs consistently improves across all tasks without additional training, resulting in an average performance enhancement ranging from **0.6** to **3.8**; (ii) These improvements hold consistently across different model sizes of 7 billion and 13 billion parameters; (iii) The efficacy extends to LLMs with varying sequence lengths, such as Vicuna-7B and its extended version, Vicuna-7B-16K, both showing improvements from **3.3** to **3.8**.

Outperform other competitive methods. We conduct a thorough comparison between Ms-PoE and other competitive methods, including Positional Interpolation (PI) [20] and Self-Extend [21], both of which modify position indices without utilizing head-wise properties. For PI, we employ the scaling ratio as the average value of our method while for Self-Extend, we set the group size as 2 with the local window size as 1024. The results presented in Table 2 consistently showcase the superiority of our approach over other baselines, demonstrating improvements of up to 3.92 and 43.72 for MDQA and Key-Value Retrieval, respectively. Such improvements might come from two primary factors. Firstly, the incorporation of head-wise properties offers a more adaptive strategy for positional modification. Secondly, our approach enhances the general context utilization ability. Notably, our approach demonstrates superiority even when the core document or key is positioned at the end of the input, surpassing other baselines with improvements ranging from 2.4 to 27.8.

https://doi.org/10.52202/079017-1943

This performance surpasses the recent work [23], which addresses the “lost-in-the-middle" effect by reordering key documents and placing them at the end of the input. When the identified core document is already located at the recent area, such method can not gain further improvements, while our approach offers a *fine-grained* strategy to improve context utilization.

Table 2: Comparsion results with other competitive methods on MDQA and Key-Value Retrival. Results are reported in accuracy.

Models	Methods	MDQA					
		1	3	5	7	10	Average
Vicuna-7B	Baseline	64.0	61.0	57.4	58.4	64.8	61.12
	PI	65.2	62.4	60.0	60.4	64.0	62.40
	Self-Extend	64.7	63.7	61.4	59.8	62.0	62.32
	Ms-PoE	**65.6**	**64.2**	**63.0**	**65.2**	**67.2**	**65.04**
Models	Methods	Key-Value Retrival					
		1	15	30	40	50	Average
Vicuna-7B	Baseline	92.0	25.8	8.0	25.4	30.0	36.24
	PI	96.4	76.4	61.4	64.6	**57.8**	67.60
	Self-Extend	88.6	63.8	76.2	59.4	42.0	66.00
	Ms-PoE	**97.0**	**83.4**	**75.0**	**86.6**	**57.8**	**79.96**

4.2 Superior Context Utilization

We assess the context utilization ability of our approaches on two tasks, including multi-document question answering (MDQA) and key-value retrieval (KV retrieval) tasks from [15]. Such tasks provide a good input structure and offers the flexibility to switch the position of crucial information, thus evaluate the context utilization ability of LLMs.

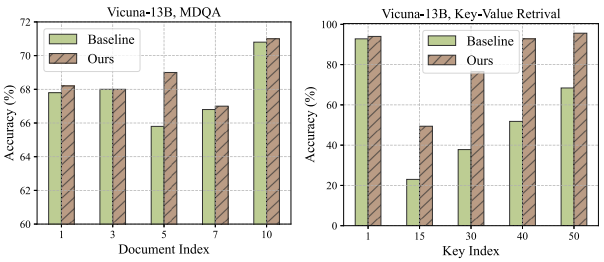

Figure 5: Comparison results for the multi-document question answering (MDQA) and key-value retrieval (KV retrieval) tasks. Each subfigure depicts the comparison when varying the position of critical information from the beginning to the end. For Vicuna-7B, please refer to Figure 1.

Experimental Setup. In the MDQA task, each input sample comprises ten documents and one question, with only one document being relevant to the question. For the KV retrieval tasks, there are 50 key-value pairs with one question querying the value of the chosen key. In both tasks, we systematically switch the important document or key-value pair from the beginning to the end and report the accuracy of the generated context. All results are averaged across 500 samples. The **Gap** accuracy metric is employed to assess the context utilization ability of LLMs, defined as the gap between the best and worst accuracy when varying the position of important information.

Main Results. As depicted in Figure 5 and 1, Ms-PoE demonstrates consistent improvement across different models, tasks and critical positions. Even when the important information exists in the sweet region (beginning and end) of the input, Ms-PoE achieves significant performance improvements ranging from 3% to 6%, highlighting its efficacy in enhancing generation quality. Moreover, the “lost-in-the-middle" issue is notably alleviated, with Ms-PoE quantitatively reducing the gap accuracy by approximately 2% to 4%, showcasing improved context utilization.

4.3 Ablation Study and More Investigation

This section conducts a further evaluation of the effectiveness of Ms-PoE by addressing the following questions: *Q1:* How does the effectiveness of Ms-PoE relate to the head-wise selection strategy of the scaling ratio? *Q2:* How does the model perform with different scaling ratios?

https://doi.org/10.52202/079017-1943

***A1:* Positional awareness metrics achieve superior performance compared to other strategies.** For a set of scaling ratios $\mathbf{r} \in R^{n_h}$, where n_h is the number of attention heads, and using scaling ratios linearly ranging from 1.2 to 1.8, we evaluate various strategies for assigning these ratios to different attention heads. These strategies include: ① `Random`, which randomly assigns the scaling ratios to each head within each layer; ② `Sequential`, performing the assignment based on the original head order; ③ `Entropy`, where we follow metrics measuring the sparsity level of attention scores [56]. Larger entropy implies less sparse attention scores, indicating the model attends to more tokens rather than just the beginning and end words, so we assign a scaling ratio near to 1, and vice versa for larger ratios. Results in Table 3 demonstrate that the proposed position-awareness effectively captures the head-wise properties of LLMs, enhancing performance when critical information is located at various positions—beginning, middle, or end. This leads to an average accuracy gain of 3.2 (65.3 v.s. 62.1).

Table 3: Ablation results of different ordering metrics. Experiments are conducted on Multi-Documents Question Answering task with the Vicuna-7B model.

Methods	Begin	Middle	End	Average
Baseline	64.0	57.4	64.8	62.1
Random	64.5	55.0	65.5	61.7
Sequential	60.5	54.5	58.5	57.8
Entropy	63.5	59.5	64.0	62.3
Position-Awareness	**65.6**	**63.0**	**67.2**	**65.3**

***A2:* Ablation study of the scaling ratios.** We first examined the effect of uniform scaling ratios across all heads on model performance. Our findings, outlined in Table 4, indicate that adjusting the scaling ratio between 0.5 and 2.5 can significantly enhance generative performance and mitigate the "lost-in-the-middle" effect by 1.0% (63.1% *v.s.* 62.1%), particularly with a ratio of 1.5. Further testing with an average ratio of 1.5 across all heads revealed that an optimal range exists between 1.2 and 1.8, leading to an additional 2.2% (65.3% *v.s.* 63.1%) accuracy improvement with our approach, Ms-PoE. Based on these results, we established these ratios as our experimental standard.

Table 4: Ablation results of the condensing ratios. Experiments are conducted on Multi-Documents Question Answering task with the Vicuna-7B model.

Scaling Ratio	Begin	Middle	End	Average
1	64.0	57.4	64.8	62.1
0.5	56.0	51.0	68.0	58.3
1.5	65.2	60.0	64.0	63.1
2	61.5	59.0	62.5	61.0
2.5	59.5	57.5	57.0	58.0
$0.8 \rightarrow 2.2$	53.5	59.5	67.5	60.2
$1 \rightarrow 2$	61.0	57.0	63.0	60.3
$1.2 \rightarrow 1.8$	**65.6**	**63.0**	**67.2**	**65.3**
$1.4 \rightarrow 1.6$	65.5	59.0	63.0	62.5

5 Conclusion

In this paper, we present a plug-and-play strategy designed to address the “lost-in-the-middle" challenge observed in LLMs. This challenge stems from the persistent bias exhibited by LLMs towards the beginning and local content within the input, leading to the neglect of crucial information in the middle. Our investigation reveals the effects of position indice rescaling and the head-wise position-awareness property, leading to the introduction of Multi-scale Positional Encoding (Ms-PoE). This approach enhances the capability of LLMs to effectively capture information in the middle of the context without the need for additional fine-tuning. Comprehensive experiments conducted on Zero-SCROLLS benchmarks, multi-document question-answering tasks, and key-value retrieval tasks confirm the effectiveness of Ms-PoE.

6 Acknowledgements

We thank Dr. Yuandong Tian for interesting discussions on this work. Z. Zhang and Z. Wang were in part supported by an Intel Gift Funding.

 https://doi.org/10.52202/079017-1943

References

[1] Chris Ré, Tri Dao, Dan Fu, and Karan Goel. Can longer sequences help take the next leap in ai?, June 2022. Accessed: 2024-01-29.

[2] Jiaqi Li, Mengmeng Wang, Zilong Zheng, and Muhan Zhang. Loogle: Can long-context language models understand long contexts? *arXiv preprint arXiv:2311.04939*, 2023.

[3] Yi Tay, Mostafa Dehghani, Samira Abnar, Yikang Shen, Dara Bahri, Philip Pham, Jinfeng Rao, Liu Yang, Sebastian Ruder, and Donald Metzler. Long range arena: A benchmark for efficient transformers. In *International Conference on Learning Representations*, 2020.

[4] Wojciech Kryściński, Nazneen Rajani, Divyansh Agarwal, Caiming Xiong, and Dragomir Radev. Booksum: A collection of datasets for long-form narrative summarization. *arXiv preprint arXiv:2105.08209*, 2021.

[5] Yusen Zhang, Ansong Ni, Ziming Mao, Chen Henry Wu, Chenguang Zhu, Budhaditya Deb, Ahmed H Awadallah, Dragomir Radev, and Rui Zhang. Summˆ n: A multi-stage summarization framework for long input dialogues and documents. *arXiv preprint arXiv:2110.10150*, 2021.

[6] Ming Zhong, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. Dialoglm: Pre-trained model for long dialogue understanding and summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11765–11773, 2022.

[7] Xueying Du, Mingwei Liu, Kaixin Wang, Hanlin Wang, Junwei Liu, Yixuan Chen, Jiayi Feng, Chaofeng Sha, Xin Peng, and Yiling Lou. Classeval: A manually-crafted benchmark for evaluating llms on class-level code generation. *arXiv preprint arXiv:2308.01861*, 2023.

[8] Qinkai Zheng, Xiao Xia, Xu Zou, Yuxiao Dong, Shan Wang, Yufei Xue, Zihan Wang, Lei Shen, Andi Wang, Yang Li, Teng Su, Zhilin Yang, and Jie Tang. Codegeex: A pre-trained model for code generation with multilingual evaluations on humaneval-x. In *KDD*, 2023.

[9] Mihaly Varadi, Stephen Anyango, Mandar Deshpande, Sreenath Nair, Cindy Natassia, Galabina Yordanova, David Yuan, Oana Stroe, Gemma Wood, Agata Laydon, et al. Alphafold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic acids research*, 50(D1):D439–D444, 2022.

[10] Shuaiwen Leon Song, Bonnie Kruft, Minjia Zhang, Conglong Li, Shiyang Chen, Chengming Zhang, Masahiro Tanaka, Xiaoxia Wu, Jeff Rasley, Ammar Ahmad Awan, et al. Deepspeed4science initiative: Enabling large-scale scientific discovery through sophisticated ai system technologies. *arXiv preprint arXiv:2310.04610*, 2023.

[11] Tri Dao. FlashAttention-2: Faster attention with better parallelism and work partitioning. 2023.

[12] Sam Ade Jacobs, Masahiro Tanaka, Chengming Zhang, Minjia Zhang, Shuaiwen Leon Song, Samyam Rajbhandari, and Yuxiong He. Deepspeed ulysses: System optimizations for enabling training of extreme long sequence transformer models. *arXiv preprint arXiv:2309.14509*, 2023.

[13] MosaicML NLP Team. Introducing mpt-7b: A new standard for open-source, commercially usable llms, 2023. Accessed: 2023-05-05.

[14] Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.

[15] Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *arXiv preprint arXiv:2307.03172*, 2023.

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

[17] Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. *arXiv preprint arXiv:2309.17453*, 2023.

[18] Zhenyu Zhang, Ying Sheng, Tianyi Zhou, Tianlong Chen, Lianmin Zheng, Ruisi Cai, Zhao Song, Yuandong Tian, Christopher Ré, Clark Barrett, et al. H _2 o: Heavy-hitter oracle for efficient generative inference of large language models. *arXiv preprint arXiv:2306.14048*, 2023.

[19] 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.

[20] 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.

[21] Hongye Jin, Xiaotian Han, Jingfeng Yang, Zhimeng Jiang, Zirui Liu, Chia-Yuan Chang, Huiyuan Chen, and Xia Hu. Llm maybe longlm: Self-extend llm context window without tuning. *arXiv preprint arXiv:2401.01325*, 2024.

[22] Weizhi Wang, Li Dong, Hao Cheng, Xiaodong Liu, Xifeng Yan, Jianfeng Gao, and Furu Wei. Augmenting language models with long-term memory. *arXiv preprint arXiv:2306.07174*, 2023.

[23] Alexander Peysakhovich and Adam Lerer. Attention sorting combats recency bias in long context language models. *arXiv preprint arXiv:2310.01427*, 2023.

[24] Yuhan Chen, Ang Lv, Ting-En Lin, Changyu Chen, Yuchuan Wu, Fei Huang, Yongbin Li, and Rui Yan. Fortify the shortest stave in attention: Enhancing context awareness of large language models for effective tool use. *arXiv preprint arXiv:2312.04455*, 2023.

[25] He Junqing, Pan Kunhao, Dong Xiaoqun, Song Zhuoyang, Liu Yibo, Liang Yuxin, Wang Hao, Sun Qianguo, Zhang Songxin, Xie Zejian, et al. Never lost in the middle: Improving large language models via attention strengthening question answering. *arXiv preprint arXiv:2311.09198*, 2023.

[26] Peng Xu, Wei Ping, Xianchao Wu, Lawrence McAfee, Chen Zhu, Zihan Liu, Sandeep Subramanian, Evelina Bakhturina, Mohammad Shoeybi, and Bryan Catanzaro. Retrieval meets long context large language models. *arXiv preprint arXiv:2310.03025*, 2023.

[27] Maor Ivgi, Uri Shaham, and Jonathan Berant. Efficient long-text understanding with short-text models. *Transactions of the Association for Computational Linguistics*, 11:284–299, 2023.

[28] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[29] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[30] Meng-Hao Guo, Cheng-Ze Lu, Qibin Hou, Zhengning Liu, Ming-Ming Cheng, and Shi-Min Hu. Segnext: Rethinking convolutional attention design for semantic segmentation. *Advances in Neural Information Processing Systems*, 35:1140–1156, 2022.

[31] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[32] Dakota Mahan, Ryan Carlow, Louis Castricato, Nathan Cooper, and Christian Laforte. Stable beluga models.

[33] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023.

[34] Uri Shaham, Maor Ivgi, Avia Efrat, Jonathan Berant, and Omer Levy. Zeroscrolls: A zero-shot benchmark for long text understanding. *arXiv preprint arXiv:2305.14196*, 2023.

 https://doi.org/10.52202/079017-1943

[35] Guanzheng Chen, Xin Li, Zaiqiao Meng, Shangsong Liang, and Lidong Bing. Clex: Continuous length extrapolation for large language models. *arXiv preprint arXiv:2310.16450*, 2023.

[36] Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. Yarn: Efficient context window extension of large language models. *arXiv preprint arXiv:2309.00071*, 2023.

[37] Matanel Oren, Michael Hassid, Yossi Adi, and Roy Schwartz. Transformers are multi-state rnns. *arXiv preprint arXiv:2401.06104*, 2024.

[38] Kaiqiang Song, Xiaoyang Wang, Sangwoo Cho, Xiaoman Pan, and Dong Yu. Zebra: Extending context window with layerwise grouped local-global attention. *arXiv preprint arXiv:2312.08618*, 2023.

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

[40] Huiqiang Jiang, Qianhui Wu, Xufang Luo, Dongsheng Li, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. Longllmlingua: Accelerating and enhancing llms in long context scenarios via prompt compression. *arXiv preprint arXiv:2310.06839*, 2023.

[41] Howard Chen, Ramakanth Pasunuru, Jason Weston, and Asli Celikyilmaz. Walking down the memory maze: Beyond context limit through interactive reading. *arXiv preprint arXiv:2310.05029*, 2023.

[42] 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.

[43] 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.

[44] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*, 2019.

[45] Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*, 2020.

[46] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[47] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

[48] Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*, 2023.

[49] 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. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

[50] Ofir Press, Noah A Smith, and Mike Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. *arXiv preprint arXiv:2108.12409*, 2021.

[51] Suyu Ge, Yunan Zhang, Liyuan Liu, Minjia Zhang, Jiawei Han, and Jianfeng Gao. Model tells you what to discard: Adaptive kv cache compression for llms. *arXiv preprint arXiv:2310.01801*, 2023.

[52] Valerii Likhosherstov, Krzysztof Choromanski, and Adrian Weller. On the expressive power of self-attention matrices. *arXiv preprint arXiv:2106.03764*, 2021.

[53] Benjamin L Edelman, Surbhi Goel, Sham Kakade, and Cyril Zhang. Inductive biases and variable creation in self-attention mechanisms. In *International Conference on Machine Learning*, pages 5793–5831. PMLR, 2022.

[54] Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Xingyu Dang, and Song Han. Awq: Activation-aware weight quantization for llm compression and acceleration. *arXiv preprint arXiv:2306.00978*, 2023.

[55] Lu Yin, You Wu, Zhenyu Zhang, Cheng-Yu Hsieh, Yaqing Wang, Yiling Jia, Mykola Pechenizkiy, Yi Liang, Zhangyang Wang, and Shiwei Liu. Outlier weighed layerwise sparsity (owl): A missing secret sauce for pruning llms to high sparsity, 2023.

[56] Yuandong Tian, Yiping Wang, Zhenyu Zhang, Beidi Chen, and Simon Du. Joma: Demystifying multilayer transformers via joint dynamics of mlp and attention. *arXiv preprint arXiv:2310.00535*, 2023.

[57] Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, et al. Longbench: A bilingual, multitask benchmark for long context understanding. *arXiv preprint arXiv:2308.14508*, 2023.

 https://doi.org/10.52202/079017-1943

A More Experiment Results

A.1 Position-Aware Attention Heads

Figure 6: Visualization of "position-aware" attention heads. Each row contains the attention pattern for the same heads when varying the key documents within the inputs.

Figure 6 illustrates the attention patterns of "position-aware" heads. Each row represents the attention pattern of the same head. As the key document is positioned from the beginning to the end, the attention peak gradually shifts, indicating robust positional awareness. It's important to note that we randomly selected 9 attention heads with these "position-aware" properties, and these results were validated with different input samples and layers.

https://doi.org/10.52202/079017-1943

Table 5: Comparison results of Ms-PoE on LongBench-EN benchmark with Llama-2-7B-Chat.

Methods	MultiFieldQA-en	LCC	GovReport	HotpotQA	Passage Count	Qasper	MultiNews	SAMSum	TriviaQA	PassageRetrieval-en	RepoBench-P	TREC	2WikiMQA	Average
Baseline	33.51	59.77	27.97	30.10	3.74	19.27	24.36	39.45	82.81	10.00	49.22	57.33	28.14	35.82
Ours	37.33	62.03	29.87	34.08	4.60	20.96	24.69	39.79	85.28	16.67	50.11	58.67	30.19	38.02

A.2 Results on LongBench-EN Benchmark

We further evaluate Ms-PoE on the LongBench-EN benchmark [57] that contains 13 tasks aims for long context understanding. Results are reported in Table 5. We can observe that Ms-PoE achieves consistent performance improvement without any finetuning.

https://doi.org/10.52202/079017-1943

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 describe sufficient details about the claims.

   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: The primary limitation of our work remains limited exploration for only RoPE based models.

   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: [NA]

Justification: This paper does not include theoretical results

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 have provided the necessary ratios for MS-PoE settings and made the relevant code publicly available in a GitHub repository.

   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?

https://doi.org/10.52202/079017-1943

Answer: [Yes]

Justification: All the models and datasets used in this paper are openly accessible on Huggingface. We made the relevant code publicly available in a GitHub repository.

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: All primary hyperparameters are presented.

   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: The majority of our results are significantly outperforming the baselines.

   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.
   - 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: All our experiments are conducted using $1 \times$ A6000 GPU.

   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: NeurIPS Code of Ethics is followed.

   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: The paper doesn't have potential positive societal impacts and negative societal impacts.

    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.

https://doi.org/10.52202/079017-1943

- 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., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: There is no risk involved.

    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: [NA]

    Justification: All assets are in public domain.

    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.
    - 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: This 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: This 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: This 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/10.52202/079017-1943