MoBA: Mixture of Block Attention for Long-Context LLMs

Technical Report

Enzhe Lu[1] **Zhejun Jiang**[1] **Jingyuan Liu**[1] **Yulun Du**[1] **Tao Jiang**[1] **Chao Hong**[1]
Shaowei Liu[1] **Weiran He**[1] **Enming Yuan**[1] **Yuzhi Wang**[1] **Zhiqi Huang**[1] **Huan Yuan**[1]
Suting Xu[1] **Xinran Xu**[1] **Guokun Lai**[1] **Yanru Chen**[1] **Huabin Zheng**[1] **Junjie Yan**[1]
Jianlin Su[1] **Yuxin Wu**[1] **Yutao Zhang**[1] **Zhilin Yang**[1]
Xinyu Zhou[1,‡] **Mingxing Zhang**[2,*] **Jiezhong Qiu**[3,‡ *†]

[1] Moonshot AI [2] Tsinghua University [3] Zhejiang Lab/Zhejiang University

2025.02

Motivation

- One key ability of large language models: Scaling the effective context length N
  - Bottleneck: the computational complexity of attention $\mathrm{O}(N^2)$

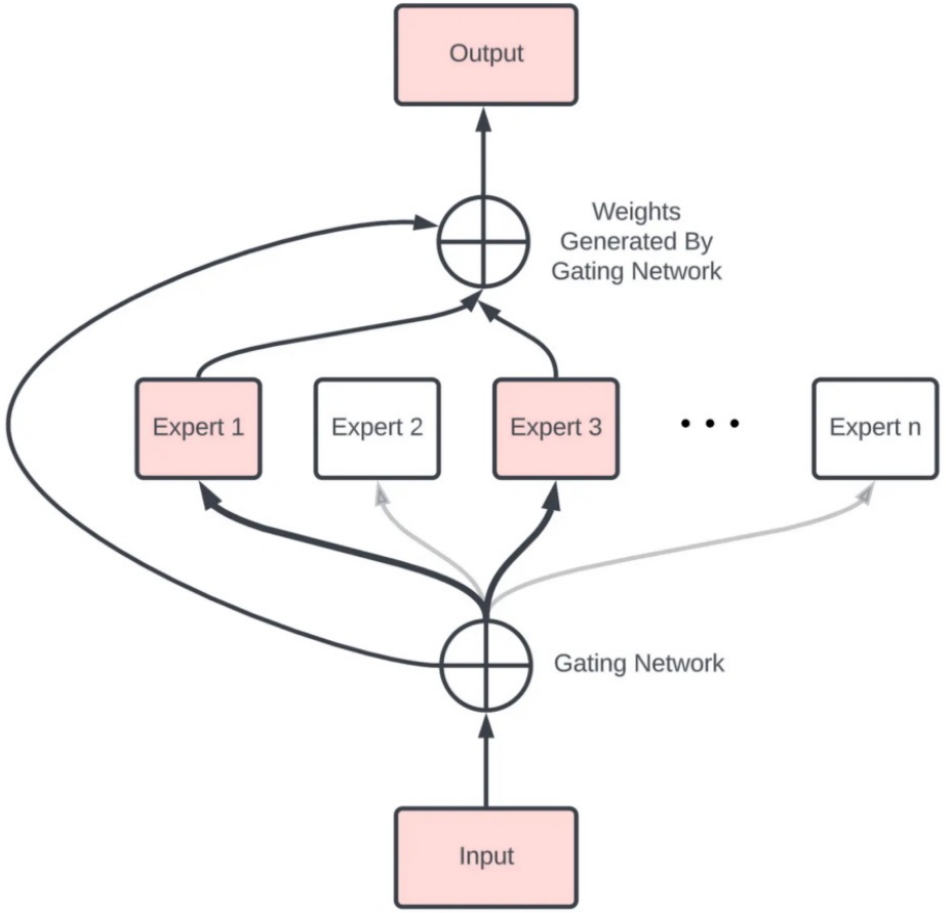
- MoE of the attention

Motivation

- MoE (for FFN)

$$MoE(x) = \Sigma_{i=1}^{K}(G_i(x)E_i(x))$$

$$G(x) = TopK(Softmax(W_g(x) + \epsilon))$$



Method

- Standard Attention in Transformer

$$\text{Attn}(\boldsymbol{q}, \boldsymbol{K}, \boldsymbol{V}) = \text{Softmax}\left(\boldsymbol{q}\boldsymbol{K}^\top\right)\boldsymbol{V}$$

- MoBA Architecture

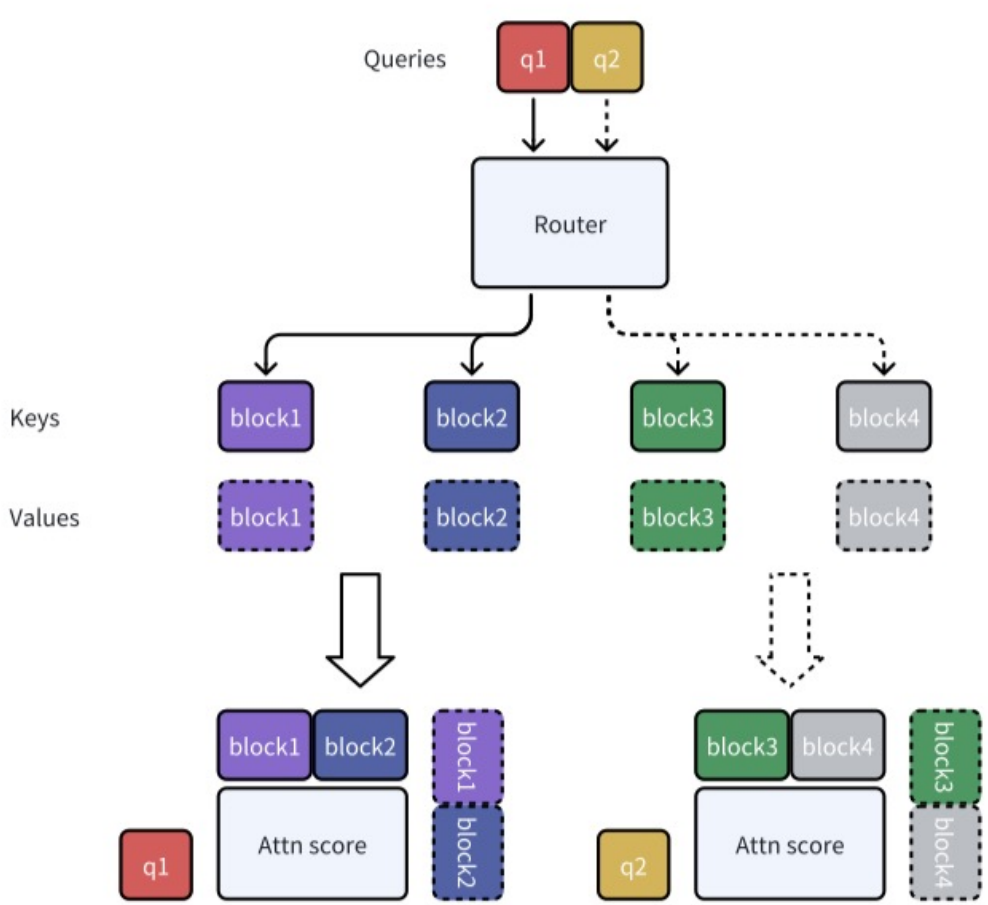
$$\text{MoBA}(\boldsymbol{q}, \boldsymbol{K}, \boldsymbol{V}) = \text{Softmax}\left(\boldsymbol{q}\boldsymbol{K}[I]^\top\right)\boldsymbol{V}[I]$$

  - N is context length
  - $I \subseteq [N]$ is the set of selected keys and values

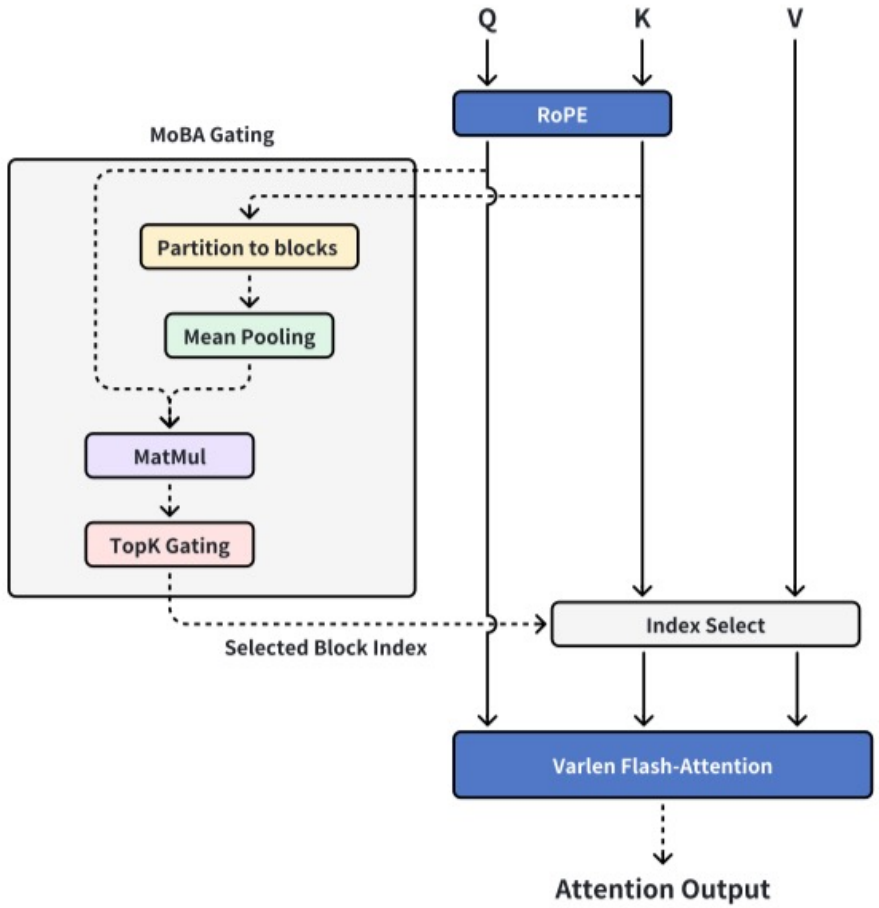
Method

- key innovation in MoBA: the block partitioning and selection strategy
  - n is the number of blocks
  - $B = \frac{N}{n}$ is block size
  - $I_i = [(i-1) \times B + 1, i \times B]$
  - gating mechanism $g_i = \begin{cases} 1 & s_i \in \text{Topk}\left(\{s_j | j \in [n]\}, k\right) \\ 0 & \text{otherwise} \end{cases} \qquad s_i = \langle \boldsymbol{q}, \text{mean\_pool}(\boldsymbol{K}[I_i]) \rangle$
  - $I = \bigcup_{g_i > 0} I_i$

Method


Queries
q1
q2
Router
Keys
block1
block2
block3
block4
Values
block1
block2
block3
block4
block1
block2
Attn score
q1
block1
block2
block3
block4
Attn score
q2
block3
block4

(a)


Q
K
V
RoPE
MoBA Gating
Partition to blocks
Mean Pooling
MatMul
TopK Gating
Selected Block Index
Index Select
Varlen Flash-Attention
Attention Output

(b)

Experiment

- MoBA vs. Flash Attention

(MoBA serves as an alternative to full attention, meaning that it does not introduce new parameters or remove existing ones.)

Experiment

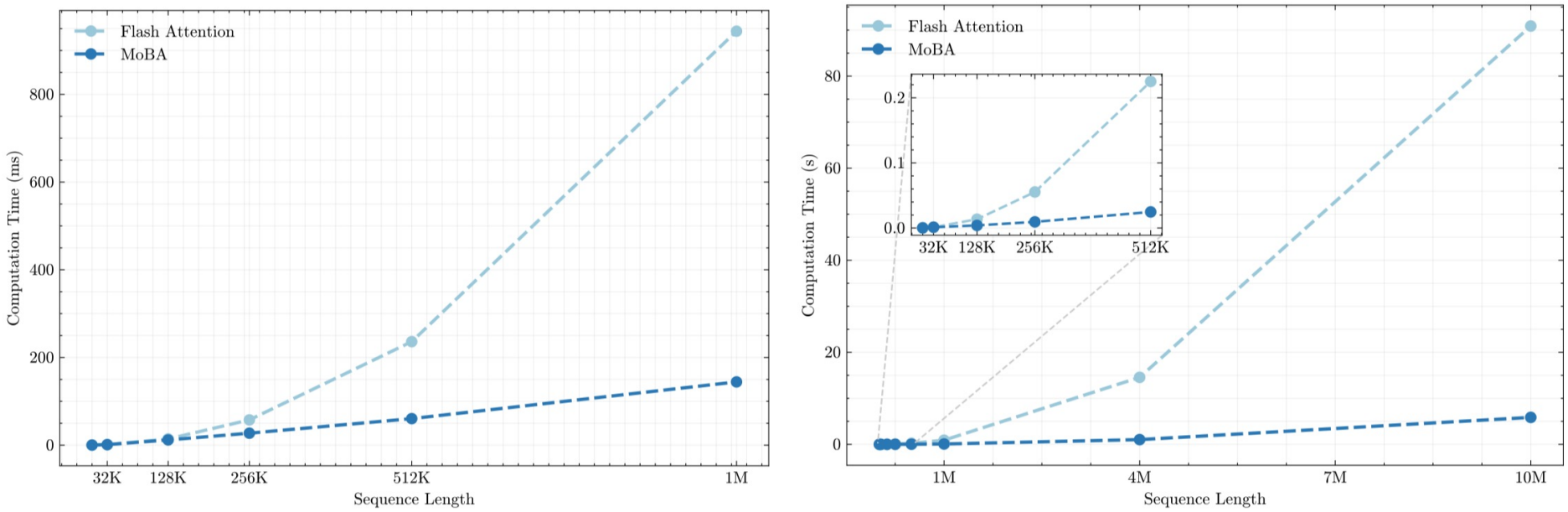
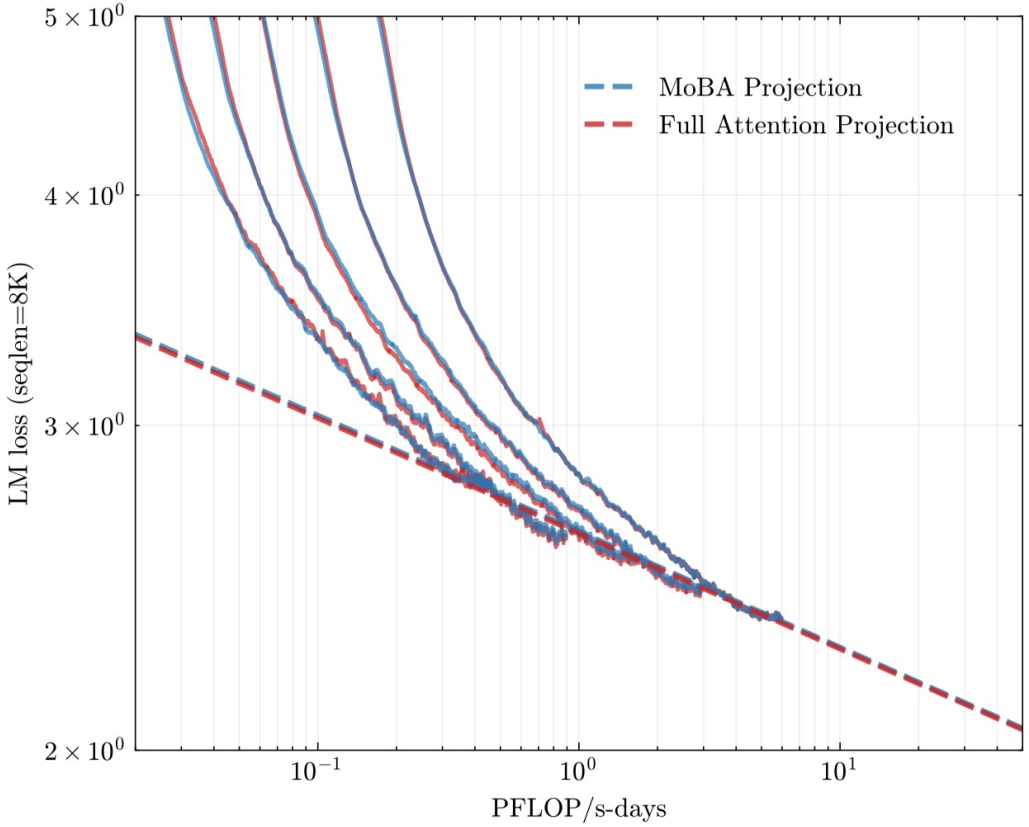


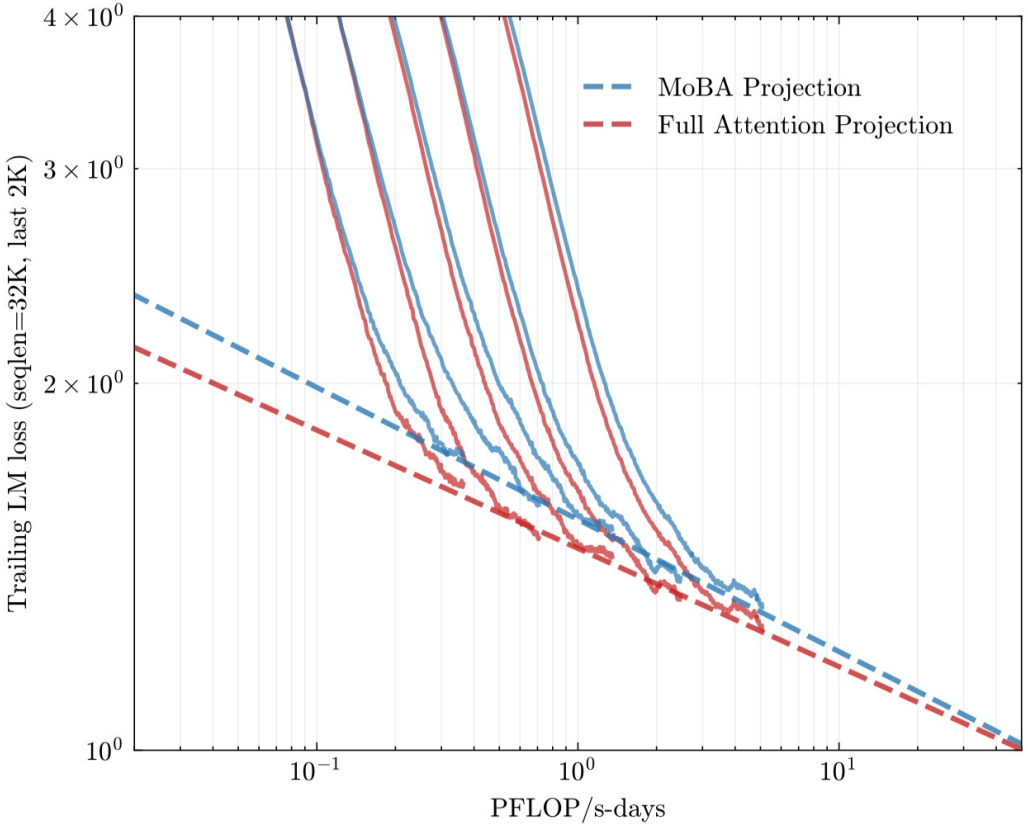
Figure 2: Efficiency of MoBA vs. full attention (implemented with Flash Attention). **(a)** 1M Model speedup evaluation: Computation time scaling of MoBA versus Flash Attention on 1M model with increasing sequence lengths (8K-1M). **(b)** Fixed Sparsity Ratio scaling: Computation time scaling comparison between MoBA and Flash Attention across increasing sequence lengths (8K-10M), maintaining a constant sparsity ratio of 95.31% (fixed 64 MoBA blocks with variance block size and fixed top-k=3).

Experiment

No-Emb Model Param	Head	Layer	Hidden	Training Token	Block size	TopK
545M	14	14	1792	10.8B	512	3
822M	16	16	2048	15.3B	512	3
1.1B	18	18	2304	20.6B	512	3
1.5B	20	20	2560	27.4B	512	3
2.1B	22	22	2816	36.9B	512	3

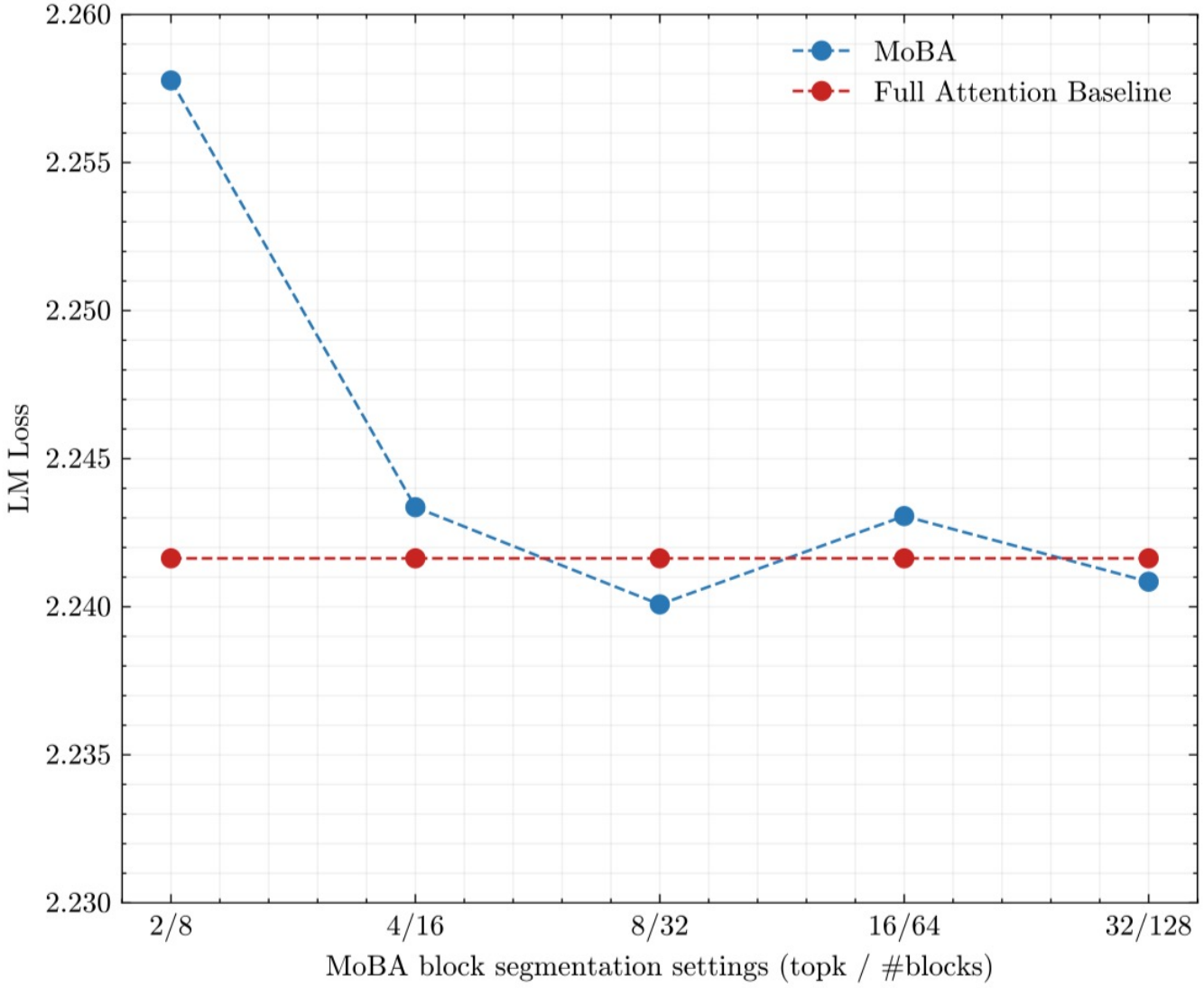


(a)

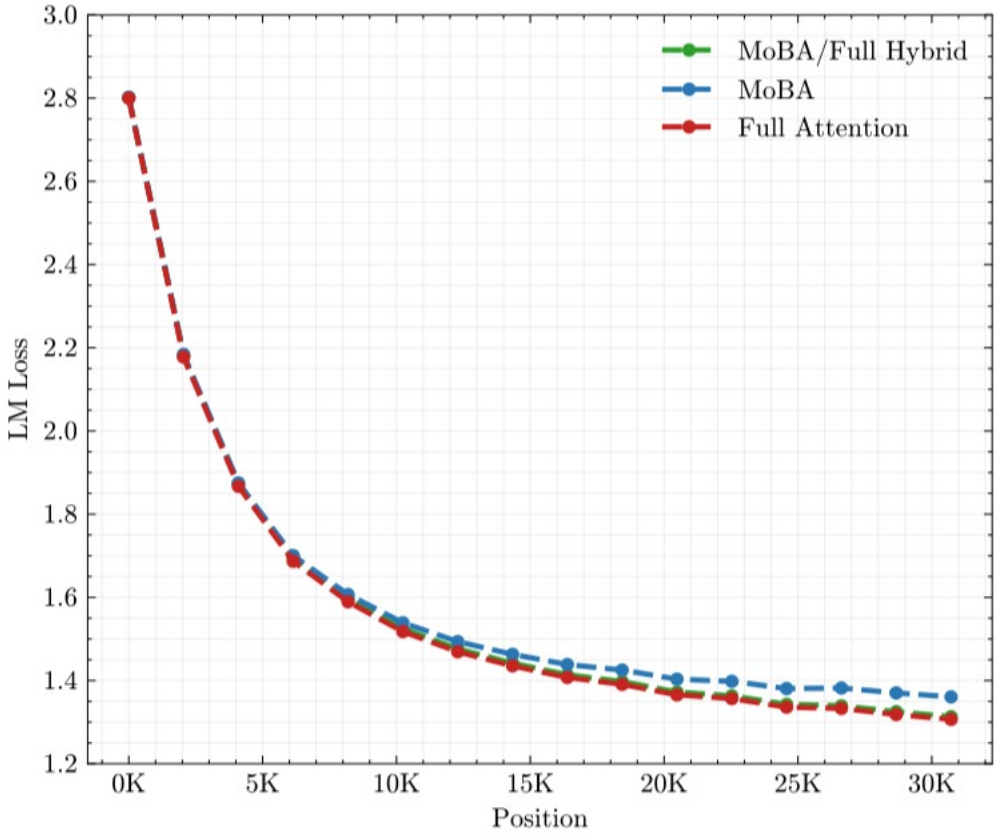


(b)

Experiment


MoBA
Full Attention Baseline
LM Loss
2.260
2.255
2.250
2.245
2.240
2.235
2.230
2/8
4/16
8/32
16/64
32/128
MoBA block segmentation settings (topk / #blocks)

Experiment


MoBA/Full Hybrid
MoBA
Full Attention
LM Loss
Position
3.0
2.8
2.6
2.4
2.2
2.0
1.8
1.6
1.4
1.2
0K
5K
10K
15K
20K
25K
30K

Summary

- Partitioning the context into blocks and employing a dynamic gating mechanism to selectively route query tokens to the most relevant KV blocks
- Offering a balanced approach between performance and efficiency

“Future work may explore further optimizations of MoBA’s block-selection strategies, investigate its application to other modalities, and study its potential for improving generalization in complex reasoning tasks”

Writing-Abstract

Scaling the effective context length is essential for advancing large language models (LLMs) toward
扩展有效的上下文长度
artificial general intelligence (AGI). However, the quadratic increase in computational complexity
计算复杂度的二次增长
inherent in traditional attention mechanisms presents a prohibitive overhead. Existing approaches either impose strongly biased structures, such as sink or window attention which are task-specific, or radically modify the attention mechanism into linear approximations, whose performance in complex reasoning tasks remains inadequately explored.

In this work, we propose a solution that adheres to the "less structure" principle, allowing the model to determine where to attend autonomously, rather than introducing predefined biases. We introduce Mixture of Block Attention (MoBA), an innovative approach that applies the principles of Mixture of
混合块注意力机制
Experts (MoE) to the attention mechanism. This novel architecture demonstrates superior
混合专家
performance on long-context tasks while offering a key advantage: the ability to seamlessly transition between full and sparse attention, enhancing efficiency without the risk of compromising
能够在全注意力和稀疏注意力之间无缝过渡
performance. MoBA has already been deployed to support Kimi's long-context requests and demonstrates significant advancements in efficient attention computation for LLMs. Our code is available at https://github.com/MoonshotAI/MoBA.

Writing-Introduction

The pursuit of artificial general intelligence (AGI) has driven the development of large language models (LLMs) to unprecedented scales, with the promise of handling complex tasks that mimic human cognition. A pivotal capability for achieving AGI is the ability to process, understand, and generate long sequences, which is essential for a wide range of applications, from historical data analysis to complex reasoning and decision-making processes. This growing demand for extended context processing can be seen not only in the popularity of long input prompt understanding, as showcased by models like Kimi (MoonshotAI 2023), Claude (Anthropic 2023) and Gemini (Reid et al. 2024), but also in recent explorations of long chain-of-thought (CoT) output capabilities in Kimi k1.5 (Team et al. 2025), DeepSeek-R1 (D. Guo et al. 2025), and OpenAI o1/o3 (Guan et al. 2024).

处理、理解和生成长序列的能力

Writing-Introduction

However, extending the sequence length in LLMs is non-trivial due to the quadratic growth in computational complexity associated with the vanilla attention mechanism (Waswani et al. 2017).

指出挑战

This challenge has spurred a wave of research aimed at improving efficiency without sacrificing performance. One prominent direction capitalizes on the inherent sparsity of attention scores. This

可能的方向：利用注意力分数的固有机制

sparsity arises both mathematically — from the softmax operation, where various sparse attention patterns have be studied (H. Jiang et al. 2024) — and biologically (Watson et al. 2025), where sparse connectivity is observed in brain regions related to memory storage.

Writing-Introduction

Existing approaches often leverage predefined structural constraints, such as sink-based (G. Xiao et al. 2023) or sliding window attention (Beltagy et al. 2020), to exploit this sparsity. While these methods can be effective, they tend to be highly task-specific, potentially hindering the model's overall generalizability. Alternatively, a range of dynamic sparse attention mechanisms, exemplified by Quest (Tang et al. 2024), Minference (H. Jiang et al. 2024), and RetrievalAttention (Di Liu et al. 2024), select subsets of tokens at inference time. Although such methods can reduce computation for long sequences, they do not substantially alleviate the intensive training costs of long-context models, making it challenging to scale LLMs efficiently to contexts on the order of millions of tokens. Another promising alternative way has recently emerged in the form of linear attention models, such as Mamba (Dao and Gu 2024), RWKV (Peng, Alcaide, et al. 2023; Peng, Goldstein, et al. 2024), and RetNet (Sun et al. 2023). These approaches replace canonical softmax-based attention with linear approximations, thereby reducing the computational overhead for long-sequence processing. However, due to the substantial differences between linear and conventional attention, adapting existing Transformer models typically incurs high conversion costs (Mercat et al. 2024; J. Wang et al. 2024; Bick et al. 2025; M. Zhang et al. 2024) or requires training entirely new models from scratch (A. Li et al. 2025). More importantly, evidence of their effectiveness in complex reasoning tasks remains limited.

预定义的结构约束

高度任务相关的

阻碍模型的整体泛化能力

Writing-Introduction

Consequently, a critical research question arises: How can we design a robust and adaptable attention architecture that retains the original Transformer framework while adhering to a “less structure” principle, allowing the model to determine where to attend without relying on predefined biases? Ideally, such an architecture would transition seamlessly between full and sparse attention modes, thus maximizing compatibility with existing pre-trained models and enabling both efficient inference and accelerated training without compromising performance.

不依赖于预定义的偏差

Thus, we introduce Mixture of Block Attention (MoBA), a novel architecture that builds upon the innovative principles of Mixture of Experts (MoE) (Shazeer et al. 2017) and applies them to the attention mechanism of the Transformer model. MoE has been used primarily in the feedforward network (FFN) layers of Transformers (Lepikhin et al. 2020; Fedus et al. 2022; Zoph et al. 2022), but MoBA pioneers its application to long context attention, allowing dynamic selection of historically relevant blocks of key and values for each query token. This approach not only enhances the efficiency of LLMs but also enables them to handle longer and more complex prompts without a proportional increase in resource consumption. MoBA addresses the computational inefficiency of traditional attention mechanisms by partitioning the context into blocks and employing a gating mechanism to selectively route query tokens to the most relevant blocks. This block sparse attention significantly reduces the computational costs, paving the way for more efficient processing of long sequences. The model’s ability to dynamically select the most informative blocks of keys leads to improved performance and efficiency, particularly beneficial for tasks involving extensive contextual information.

允许动态选择每个query的key和value的历史相关块

Writing-Introduction

In this paper, we detail the architecture of MoBA, firstly its block partitioning and routing strategy, and secondly its computational efficiency compared to traditional attention mechanisms. We further present experimental results that demonstrate MoBA's superior performance in tasks requiring the processing of long sequences. Our work contributes a novel approach to efficient attention computation, pushing the boundaries of what is achievable with LLMs in handling complex and lengthy inputs.