MOBA：长篇文化LLM的阻止注意力的混合

技术报告

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] **Neo Y. 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

ABSTRACT

扩展有效上下文长度对于将大型语言模型（LLM）推向人工通用情报（AGI）至关重要。然而，传统注意力机制中计算复杂性的二次增加呈现出一个高度的开销。现有的近距离施加了强烈偏见的结构，例如，特定于任务的下沉或窗户注意力，将注意力机制修改为线性近似值，其在com-plex推理任务中的性能仍然不足。对于“较少的结构”原则，允许模型确定自主参加的何处，而不是引入预定义的偏见。我们介绍了阻止注意力（MOBA）的混合物，这是一种创新的方法，将专家（MOE）的原理应用于注意机制。这种新颖的体系结构表明了长篇文化任务上的su-prorior绩效，同时提供了一个关键优势：在全面和稀疏注意力之间无缝地进行无缝延伸的能力，提高效率而不会受到损害的风险。 MOBA已经被部署以支持Kimi的长篇文章请求，并在LLMS有效注意计算方面的魔鬼构成重大进步。我们的代码可用https://github.com/moonshotai/moba。

1 简介

对人工智能（AGI）的追求驱动了大型语言模型（LLMS）的发展量表，并有望处理模仿人类认知的复杂任务。实现AGI的关键能力是处理，理解和生成长序列的能力，这对于从历史数据分析到复杂的推理和决策过程至关重要。这种扩展上下文处理的种类不仅可以在长期输入及时理解的流行中看到，如Kimi（Moonshotai 2023），Claude（Anthropic 2023）和Gemini（Reid等人（Reid etal。2024））所展示的那样，还可以看到。 Kimi K1.5中长期思考（COT）输出功能的探索（Team等。 2025），DeepSeek-R1（D. Guo等，2025）和Openai O1/O3（Guan等，2024）。

然而，由于与香草注意机制相关的计算复杂性的二次增长，LLMS中的序列长度是非平凡的（Waswani等，2017）。这项挑战激发了一波研究，旨在提高效率而不牺牲绩效。一个突出的方向利用了注意力评分的稀疏性。这种稀疏性在数学上都是出于软马克斯的操作而产生的，从而研究了一种稀疏的注意力模式（H. Jiang等，2024），也可以从生物学上进行（Watson等，2025），在与记忆存储有关的大脑区域中观察到Wheresparse的连接性。

[*]zhang_mingxing@mail.tsinghua.edu.cn
[†‡]Co-corresponding authors. Xinyu Zhou (zhouxinyu@moonshot.cn), Jiezhong Qiu (jiezhongqiu@outlook.com)

现有方法通常利用预定义的结构约束，例如基于水槽的（G. Xiao等，2023）或滑动窗户的注意力（Beltagy等，2020），以利用这种稀疏性。尽管这些方法可能是有效的，但它们倾向于特定于任务，这可能会阻碍该模型的总体普遍性。另外，通过Quest（Tang等，2024年），终结（H. Jiang等，2024）和重新进行的一系列动力学注意机制（Di Liu等，2024年），选择了Tokens of Potkens of Presprence t.时间。尽管这种方法可以降低长序列的缩小，但它们并不能大大减轻长篇小说模型的强化培训成本，从而使LLM有效地缩放到有关数百万个令牌的上下文的挑战。最近以线性注意模型的形式出现了另一种有希望的方向方法，例如Mamba（Dao和Gu 2024），RWKV（Peng，Alcaide等，2023; Peng，Peng，Goldstein等，2024年）和Retnet（Sun et et and Sun et et necte Al。这些接近典范的典范基于线性近似的关注，从而减少了长期处理的计算间接费用。但是，由于线性和常规关注之间存在实质性差异，适应现有变压器模型通常会产生高转换成本（Mercat等，2024; J. Wang等，2024; Bick etal。2025; M. Zhang etal。2024; ）或需要从头开始培训全新的模型（A. Li等，2025）。更重要的是，它们在复杂的推理任务中有效性的证据仍然有限。

因此，出现了一个批判性的研究问题：我们如何设计一种强大而适应性的注意力结构，可以在遵守“较少结构”原则的同时启用原始的变压器框架，从而允许模型在不依赖预定义偏见的情况下进行的模型在哪里参加？理想情况下，这样的架构将在全面和稀疏的注意模式之间无意义地过渡，从而最大程度地提高与现有预训练模型的兼容性，并实现有效的推理和加速训练，而不会损害性能。

因此，我们引入了块关注（MOBA）的混合物，这是一种基于专家（MOE）的创新原理（Shazeer等人，2017年）的创新原理，并将其应用于TransformerModel的注意机制。 MOE主要用于变压器的前馈网络（FFN）层（Lepikhin等，2020； Fedus等，2022； Zoph等，2022），但是MOBA先驱者将其在长篇小说中的应用中应用，从而使历史上相关的动态启动每个查询令牌的密钥和值块。这种方法不仅提高了LLM的效率，而且还使他们能够处理更长，更复杂的提示，而无需成比例的增加资源消耗。 MOBA解决了传统注意机制的计算效率低下，将上下文分配到块中，并采用门控机制选择性地将查询令牌路由到mostretlevant块。这种障碍稀疏的注意力大大降低了计算成本，为长序列处理更有效的处理铺平了道路。该模型动态选择最有用的障碍的能力会提高性能和效率，这对于涉及广泛的上下文信息的任务尤其有益。

在本文中，我们详细介绍了MOBA的架构，首先是其块分配和路由策略，其次与传统的注意机制相比，其计算效率。我们进一步提出了实验结果，可以证明MOBA在需要处理长序列的任务中的出色表现。我们的工作为有效的注意力计算做出了贡献的方法，从而突破了LLM在处理复杂和冗长输入中可实现的界限。

2 方法

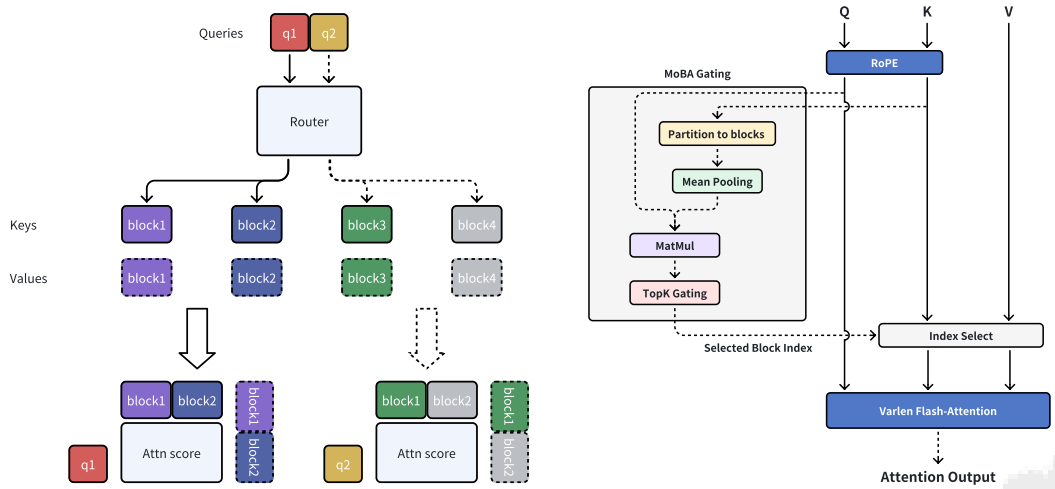
在这项工作中，我们介绍了一种新颖的架构，称为“块注意力（MOBA）”的混合物，该结构通过动态选择历史片段（块）进行注意计算来扩展变压器模型的CA-PIS能力。 （MOE）和稀疏的注意力。前一种技术已在变压器体系结构内的前馈网络（FFN）层中进行了预先应用，而后者在缩放变压器方面广泛采用以处理长上下文。我们的方法在将注意力集团应用于注意机制本身方面具有创新性，从而可以更有效地处理长序列。

2.1预赛：变压器中的标准注意力

我们首先重新审视变形金刚中的标准关注。为简单起见，我们重新审视单个查询tokenq∈R1×d分别参与n键和值代币（表示k，v∈Rn×d）。标准注意力归结为：

$$\text{attn}(q, k, v) = \text{softmax}\, qk^{\top}v, \quad (1) \tag{1}$$

其中D表示单个注意力头的维度。为了清楚起见，我们专注于单头情景。对多头注意的扩展涉及将多个此类单头注意操作的输出串联。



(a) (b)

图1：阻止注意的混合物（MOBA）的说明。 （a）MOBA的运行示例； （b）将moba整合到闪烁的注意力中。

2.2 MoBA Architecture

与标准关注不同的是，每个查询代币都在整个上下文中都参与，MOBA可以使每个查询都只能参与键和值的子集：

$$\text{moba（q，k，v）= softmax qk [i]Tv [i]，（2）} \top \Big) \boldsymbol{V}[I], \tag{2}$$

其中i⊆[n]是选定的键和值的集合。

MOBA的关键创新是块分配和选择策略。我们将长度n块的完整上下文划分，其中每个块代表后续令牌的子集。如果没有损失一般性，我们假设上下文长度n可以由块n的数量排除。我们进一步表示b = nn是块大小andi i = [（i -1）×b + 1，i×b]（3）

成为第三块的范围。通过应用MOE的TOP-K门控机制，我们可以启用每个查询，将每个查询都集中在不同块的子集上，而不是整个上下文：

$$I = \bigcup_{g_i > 0} I_i. \tag{4}$$

该模型采用门控机制，如等式4中的G i，以选择每个查询令牌的最相关块。MOBAGate首先计算亲和力得分S I测量查询Q和i-th块之间的相关性，并应用顶部 - top- k在所有街区中的门控。更正式地，I-th Block G I的门值是由

$$g_i = \begin{cases} 1 & s_i \in \text{Topk}\left(\{s_j | j \in [n]\}, k\right) \\ 0 & \text{otherwise} \end{cases}, \tag{5}$$

其中topk（·，k）表示每个块计算的亲和力分数中包含k个最高分数的集合。在这项工作中，得分s是由Q和沿序列的k [i]平均池之间的内部产物计算的：s i =⟨q，平均池（k [i i]）⟩（6）

-

一个运行示例。我们在图1A上提供了一个MOBA的运行示例，其中有两个查询令牌和四个KV块。路由器（门控网络）动态选择每个查询的前两个块。在图1A中进行了回答，第一个查询分配给第一个和第二个块，而第二查询分配给第三和第四块。

重要的是要在自回归语言模型中保持因果关系，因为它们通过以前的令牌基于下一步的预测生成文本。这个顺序的生成过程可确保令牌不能影响它之前的令牌，从而保留因果关系。 MOBA通过两个特定设计保存因果关系：

因果关系：不关注未来的障碍。 MOBA确保无法将查询令牌路由到将来的任何块。通过将注意力范围限制在当前和过去的块上，MOBA遵守LanguageModeling的自回归性质。更正式地，将pos（q）表示为查询q的位置索引，我们将s i =-∞和g i = 0设置为i blocks i，以使pos（q）

当前的关注和因果掩盖。我们将“当前块”定义为包含QueryToken本身的块。与当前区块的路由也可能违反因果关系，因为无意间在整个区块中平均汇总包含未来代币的信息。为了解决这个问题，我们强制执行每个令牌必须将其各自的当前块路由并在当前块注意力期间施加因果面具。该策略不仅避免从后续令牌中泄漏任何信息，还鼓励人们注意当地环境。更主张，我们为块i设置了g i = 1，其中查询令牌pos的位置（q）在间隔i i内。从专家的混合物（MOE）的角度来看，MOBA中的当前关注类似于现代MOE架构共享专家的作用（Dai等，2024; A. Yang等，2024），其中静态路由规则是添加专家选择时。

接下来，我们将讨论MOBA的一些其他关键设计选择，例如其块细分策略和MOBA混合物以及全部关注。

细粒块细分。精细颗粒专家分割在改善模态性能方面的积极影响已得到充分记录在专家（MOE）文献中（Dai等，2024; A. Yang et et al.2024）。在这项工作中，我们探讨了应用类似的细粒分割技术MOBA的潜在优势。 MOE启发的MOBA沿上下文长度的维度而不是FFNInterMediate隐藏维度进行分割。因此，我们的调查旨在确定当Wepartition将上下文分解成细晶粒的区块时，MOBA是否也可以受益。在第3.1节中可以找到更多的实验结果。

MOBA的混合和全部关注。 MOBA被设计为全部关注的替代品，可维持参数的Samenumber，而无需任何添加或减法。此功能激发了我们在全面关注和MOBA之间进行平稳的过渡。具体而言，在初始化阶段，每个注意力层都有选择的选择或MOBA，并且在必要时可以在训练过程中动态更改此选择。在先前的工作中已经研究了对滑动窗户关注的全面关注的类似想法（X. Zhang等，2024）。实验性结果可以在第3.2节中找到。

与滑动的窗户注意力和注意力下沉相比。滑动窗户的注意（SWA）和关注点，sinkare两个流行的稀疏注意体系结构。我们证明，两者都可以看作是MOBA的特殊情况。对于滑动窗户的关注（Beltagy等，2020），每个查询令牌都只能照顾其附近的令牌。这可以解释为具有门控网络的MOBA变体，该网络不断选择最新的块。同样，注意下沉（G. Xiao等，2023），每个查询令牌都可以参与初始令牌和mostrecent代币的组合，可以看作是MOBA的变体，带有门控网络，始终选择初始和更额外的obs网络块。上面的讨论表明，与滑动窗户注意力和关注点相比，MOBA具有更强的表现力。此外，它表明，MOBA可以灵活地近似许多静态稀疏注意架构，而不是结合特定的门控网络。

总体而言，MOBA的注意机制使该模型可以自适应地专注于上下文中最有用的块。这对于涉及长文档或序列的任务尤其有益，在整个上下文中可能不必要且计算昂贵。 MOBA有选择地参加Torelevant障碍的能力可以使信息更加细微，有效地处理。

2.3 实施

我们通过结合Flashatten-Tion（Dao，D。Fu等，2022）和MOE（Rajbhandari等，2022）的优化技术来提供MOBA的高性能实现。图2证明了MOBA的高效率，而我们将有关效率和可伸缩性的详细实验推荐为第3.4节。我们的实施是一个有意义的主要步骤：

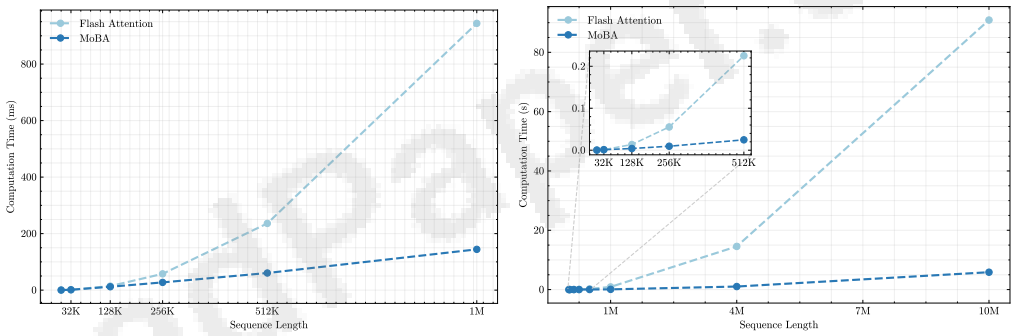
· 根据门控网络和因果面具，确定查询令牌对KV块的分配。

· 根据分配的KV块安排查询令牌的订购。

· 计算每个KV块的注意力输出，并分配给其的查询令牌。可以以不同的长度来优化此步骤。

算法1 MOBA（关注块的混合）实施

要求：查询，键和值矩阵q，k，v∈Rn×h×d; MOBA超参数（B尺寸B和TOP-K）；手D表示注意力头和头尺寸的数量。另外，表示n = n/b为blocks.1：//将kv分为blocks2：{ki，ṽi} = split blocks（k，k，v，b），其中ki，ṽi，ṽI∈Rbb×h×d，i ∈[n] 3：//计算动态块选择的门控评分4：k =平均池（k，b）∈Rn×h×d5：s = qkT ∈Rn×h×n6：//选择有因果约束的块（不关注未来块）7：M =创建因果面具（n，n）8：g = topk（s + m，k）9：//计算效率的注意模式10：q s，ks，ṽs=获取自我attn块（q，k，ṽ）11：q m，km，ṽm= index select select oba attn块（q，k，ṽ，g）12：//分别计算关注13：o s = flash fastion varlen（q s，ks，ṽs，Causal = true）14：o m = flash faster注意varlen（q m，km，km，km，km，ṽM，因果= false）15：//将结果与在线SoftMax16：o =结合在线SoftMax（O S，O M）17：返回o



(a) (b)

图2：MOBA与全部关注的效率（以闪光灯的注意实施）。（a）1M模型加速评估：MOBA与闪光注意的计算时间缩放在1M模型上具有增加序列长度（8K-1M）。（b）固定的稀疏性比缩放：MOBA和Flash注意力横幅之间的计算时间缩放比较增加了序列长度（8K-10M），保持恒定的稀疏比为95.31%（固定64 MOBA Blockswith Variancation Block Block大小和固定的TOP-K = 3）。

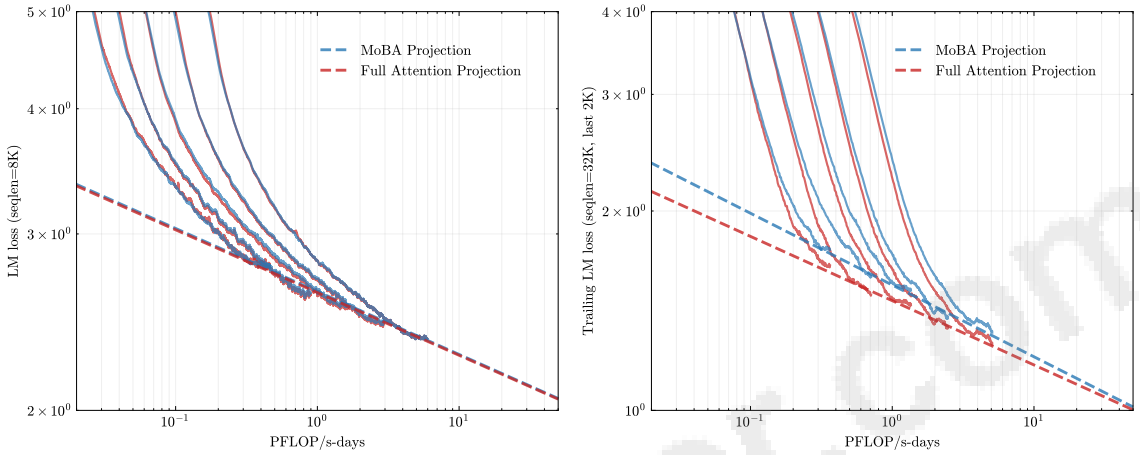
· 重新安排注意力输出回到其原始订购。

· 使用在线SoftMax（即瓷砖）组合相应的注意输出，因为查询令牌可能会访问其电流块和多个历史kV块。

该算法工作流程在算法1中正式化，并在图1B中进行了可视化，说明了如何根据MOE和Flashingention进行MOBA。首先，将KV矩阵划分为块（第1-2行）。接下来，根据等式6计算门控评分，该等式6衡量查询令牌和kvblocks之间的相关性（第3-7行）。在门控分数（与因果面罩一起使用）上，将TOP-K运算符应用，从而导致SPARSECORY-KV-BLOCK映射矩阵G表示查询分配给KV块（第8行）。然后，基于查询到kV块映射的QueryTokens是安排的，并计算出块的注意输出（Line9-12）。值得注意的是，对历史块（第11和14行）的关注以及当前的关注（第10和13行）是单独计算的，因为需要在当前的障碍注意力中保持其他因果关系。最后，将输出的输出重新排列回原始订购，并与在线软马克斯（第16行）相结合（Milakovetal。2018; H. Liu etal。2023）。

Model Param	Head	Layer	Hidden	Training Token	Block size	TopK
568M	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

Table 1: Configuration of Scaling Law Experiments



(a) (b)

L(C)	MoBA	Full
LM loss (seqlen=8K)	$2.625 \times C^{-0.063}$	$2.622 \times C^{-0.063}$
Trailing LM loss (seqlen=32K, last 2K)	$1.546 \times C^{-0.108}$	$1.464 \times C^{-0.097}$

(c)

图3：缩放定律比较MOBA和完全关注。（a）验证集的LM损失（seqlen = 8k）；（b）验证集的LM损失（seqlen = 32k，最后1k代币）；（c）合适的缩放法律曲线。

3 实验

3.1缩放定律实验和消融研究

在本节中，我们进行缩放法实验和消融研究，以验证MOBA的一些关键设计选择。

可伸缩性W.R.T. LM损失。为了评估MOBA的有效性，我们通过比较了使用完全关注或MOBA训练的语言模型的验证丢失来执行缩放定律实验。遵循Chinchilla Scalinglaw（Hoffmann等，2022），我们训练五种不同尺寸的语言模型，并以足够数量的训练来确保每个模型都能达到其最佳训练。表1中发现的缩放法实验的详细配置。MOBA和全注意模型均以8K的序列长度进行训练。对于Mobamodels，我们将块大小设置为512，然后选择前3个块以进行注意，从而导致稀疏的注意力模式，稀疏性高达1-512×38192 = 81.25%3。特别是，MOBA可以替代全神贯注，这意味着它不会引入新参数或删除现有参数。这种设计简化了我们的比较过程，因为所有实验的唯一差异都在于注意模块，而所有其他超参数（包括三重速率和批处理尺寸）仍然保持恒定。如图3A所示，MOBA和全面注意的验证损失曲线显示出非常相似的缩放趋势。具体而言，这两种注意力机制之间的验证损失差异在1E -3的范围内保持一致。这表明MOBA达到的缩放性能与完全关注的缩放性能相当，尽管其注意力稀少，而稀疏性高达75%。

3由于我们设置了top-k = 3，因此每个查询令牌最多都可以参与2个历史记录块和当前块。

长上下文可扩展性。但是，LM损失可能会因数据长度分布而偏斜（An等，2024），该分布通常由短序列主导。为了充分评估MOBA的长期文字能力，我们评估了LM损失的尾随令牌（简而最大序列长度以避免可能出现短序列的偏见。有关尾随令牌缩放的详细讨论，请参见附录A.1

这些指标为模型生成序列的最终部分的能力提供了见解，这对于涉及长上下文理解的任务可以提供信息。因此，我们通过将最大序列长度从8K增加到32K来采用改进的实验集。这种调整导致对MOBA的注意力更大，达到1-512×332768 = 95.31%的稀疏度水平。如图3B所示，尽管Mobaexhib在所有五个实验中的全部关注度相比，LM的最后一个块损失略高，但损失差距在狭窄方面缩小。该实验意味着MOBA的长篇文化可伸缩性。

关于细粒块分割的消融研究。我们进一步消除了MOBA的块状粒度。我们使用具有32K上下文长度的1.5B参数模型进行一系列实验。调整块和TOP-K的超参数以保持一致的注意力稀疏度。具体而言，我们将32k ContextInto 8、16、32、64和128个区块划分，并相应地选择2、4、8、16和32个块，以确保在这些配置中的注意力为75%。如图4所示，MOBA的性能受到块状性的显着影响。具体而言，在最粒度的设置（从8个中选择2个窗口）和具有更精细的粒度的设置之间的性能差为1E-2。这些发现表明，细粒细分表现出是一种通用技术，可以增强包括MOBA在内的MoE家族中模型的性能。

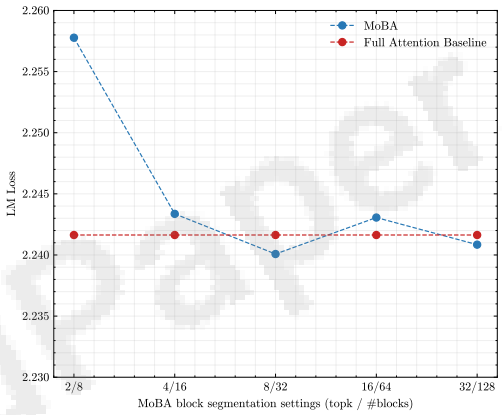


图4：细粒块分割。 LM验证集v.s的损失具有不同块粒度的MOBA。

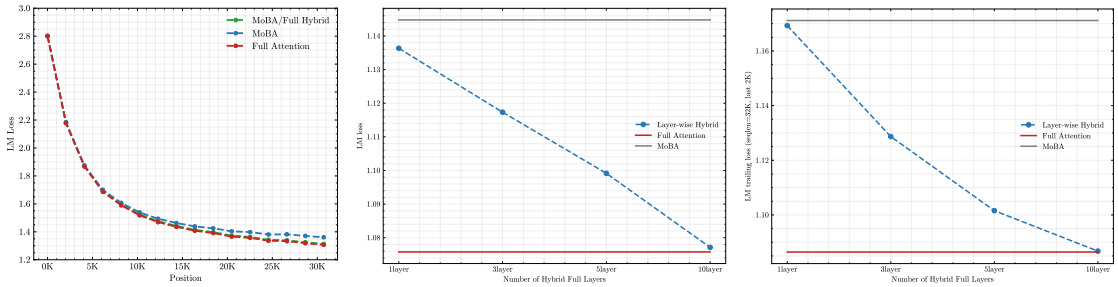
3.2 MOBA的混合和全部关注

如第2节所述，我们将MOBA设计为全部关注的灵活替代品，以便它可以轻松地从开心/通过最小的开销进行切换/全部关注，并实现可比的长篇小说性能。在本节中，Wefirst显示了全面关注和MOBA之间的无缝过渡，可以作为有效的长篇文化预训练的解决方案。然后，我们讨论了层面上的混合策略，主要是针对监督微调（SFT）的性能。

MOBA/完整的混合动力培训。我们在30B令牌上训练三个模型，每个模型具有1.5B参数，上下文长度为32K令牌。对于MOBA的超参数，将块大小设置为2048，而Top-K参数为Setto 3。详细的培训配方如下：

· MOBA/FULL HYBRID：使用两个阶段配方对此模型进行训练。在第一阶段，MOBA用于训练90%的令牌。在第二阶段，该模型对剩余的10%令牌的全部关注。 · 全部关注：在整个培训中，使用全面关注对此模型进行了训练。 · MOBA：此模型仅使用MOBA培训。

我们通过位置语言模型（LM）损失评估了它们的长期性能，这是一个精细的元素，可以评估序列内每个位置的LM损失。与通过所有位置的LM损失计算得出的香草LM损失不同，位置LM损失分别分别分解了每个位置的损失。 2024年），他注意到



(a) (b) (c)

图5：MOBA的混合和全部关注。（a）由MOBA的位置LM损失，全部关注和MOBA/Full Hybrid培训；（b）sft lm损失W.R.T层杂种中的全部注意层的数量；（c）sft tailding lmloss（seqlen = 32k，最后2k）w.r.t thease Hybrid中的全部注意力层的数量。

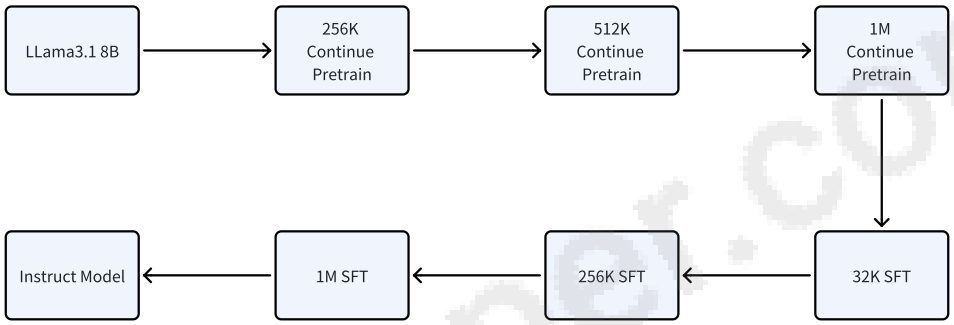


图6：持续的预训练和SFT食谱。

位置LM损失遵循相对于上下文长度的幂律趋势。如图5A所示，MOBA独立式分组会导致落后令牌的较高位置损失。重要的是，我们的MOBA/FULL HYBRID食谱达到了与完全关注的食物相同的差异。该结果凸显了MOBA/Full Hybrid TrainingRecipe在平衡训练效率与模型性能之间的有效性。更有趣的是，我们没有观察到在MOBA和完全注意力之间的切换过程中明显的尖峰，这再次证明了Moba的灵活性和鲁棒性。

层的混合动力。 MOBA的这种灵活性鼓励我们深入研究更复杂的策略 - thelayer的MOBA杂种和全部关注。我们特别着眼于其应用程序，以调查该策略，以使受监督的微调（SFT）。调查这种策略的动机源于我们的观察结果，即摩ba有时会在SFT期间导致次优性能，如图5B所示。我们推测，这可能归因于SFT中使用的丢失掩盖 - 及时令牌通常被排除在损失计算的SFT之外，这可能会对像MOBA这样的稀疏注意方法构成稀疏的梯度挑战。因为它在整个eContext中都会根据未掩盖的令牌计算的梯度的反向传播。为了解决这个问题，我们提出了一种混合方法 - 将最后几个变压器层从Moba转换为全部关注，而其余层则继续采用MOBA。如图5B和图5C所示，此策略可以显着减少SFT损失。

3.3大语言建模评估

我们在各种下游任务中对MOBA进行了彻底的评估，与全注意模型相比，评估了其绩效。为了易于验证，我们的实验始于Llama 3.18B基本模型，该模型被用作长篇文化预训练的起点。该模型称为Llama-8B-1M-MOBA，最初以128K令牌的上下文长度进行训练，并且在持续的预训练期间，我们逐渐增加了上下文长度TO256K，512K和1M令牌。为了缓解这种过渡，我们在256K持续训练阶段开始时使用位置插值仪（S. Chen等，2023）。这项技术使我们能够扩展

Benchmark	Llama-8B-1M-MoBA	Llama-8B-1M-Full
AGIEval [0-shot]	0.5144	**0.5146**
BBH [3-shot]	0.6573	**0.6589**
CEval [5-shot]	**0.6273**	0.6165
GSM8K [5-shot]	**0.7278**	0.7142
HellaSWAG [0-shot]	0.8262	**0.8279**
Loogle [0-shot]	**0.4209**	0.4016
Competition Math [0-shot]	0.4254	**0.4324**
MBPP [3-shot]	**0.5380**	0.5320
MBPP Sanitized [0-shot]	**0.6926**	0.6615
MMLU [0-shot]	0.4903	**0.4904**
MMLU Pro [5-shot][CoT]	0.4295	**0.4328**
OpenAI HumanEval [0-shot][pass@1]	0.6951	**0.7012**
SimpleQA [0-shot]	0.0465	**0.0492**
TriviaQA [0-shot]	**0.5673**	0.5667
LongBench @32K [0-shot]	**0.4828**	0.4821
RULER @128K [0-shot]	0.7818	**0.7849**

Table 2: Performance comparison between MoBA and full Attention across different evaluation benchmarks.

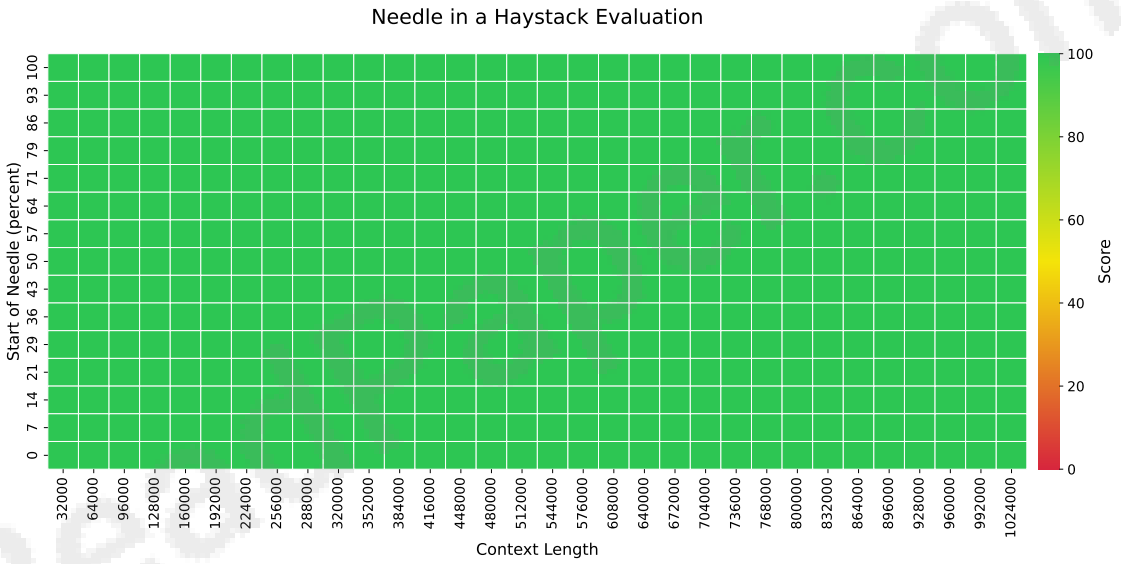


图7：在干草堆基准中的针头上的骆驼-8b-1m-moba（最高上下文长度）。

有效上下文长度从128K令牌到1M令牌。完成1m连续预训练后，Mobais激活了100B令牌。我们将块大小设置为4096，将TOUP-K参数设置为12，导致高达1-4096×12
= 95.31%。为了保留一些全面的注意力，我们采用了层面的混合材料 - 最后三层保持着全部关注，而其他29个全部注意力层转向MOBA。用于监督的微调，我们遵循类似的策略，逐渐增加上下文长度从32K到1。基线全部注意模型（称为Llama-8b-1m-full）也遵循类似的训练策略，如显示的6，唯一的区别是使用在整个过程中都受到全部关注。这种方法使我们可以直接将MOBA的性能与同等训练条件下的全部注意力模型的性能进行比较。

该评估是在几个广泛使用的长篇文章基准上进行的。特别是，在所有评估任务中，MOBA仅用于预填充，而我们在生成期间转移到更好的性能。如图2所示，Llama-8b-1m-Moba表现出与Llama-8b-1m-Full相当的性能。特别值得注意的是，在最长的基准中，统治者的稀疏度高达1-4096×12128K = 62.5%，Llama-8b-1m-Moba与Llama-8B-1M-Full的性能匹配，与Llama-8b-1m-Moba相匹配得分为0.7818，而0.7849。对于最高1M令牌的上下文长度，我们使用传统的针线素（Haystack Bench Markec）评估了该模型。如图7所示，Llama-8b-1m-Moba表现出令人满意的性能，即使上下文长度为100万个令牌。

3.4效率和可扩展性

上述实验结果表明，MOBA不仅在语言模型方面，而且在现实世界中都能达到可比的性能。为了进一步研究其效率，我们比较了在第3.3节中训练的两个模型（Llama-8b-1m-Moba和Llama-8b-1m-full）中的theattention层的前进时间。我们将注意力层的集中精力集中在两个模型中的所有其他层（例如FFN）中都具有相同的拖曳。如图2A所示，在所有上下文长度上，MOBA比全部关注更有效，表明了次级计算的次数。特别是，在预填充1M令牌时，它的加速度比率高达6.5倍。

我们还通过将上下文长度逐渐提高到1000万个令牌来探索MOBA的长度可伸缩性。稳定的注意力稀疏性，我们将固定的顶部K值和MOBA块的数量保持固定，同时成比例地缩小块大小。为了达到10m上下文长度，我们将张量并行性（Shoeybi etal。2019）扩展到了查询头级，具体来说，我们在分布式查询头上广播钥匙和值张量，从而有效地解决了GPU记忆限制，同时保持了计算效率。如图2B所示，与标准闪光注意力相比，当缩放到更长的序列时，MobAdjext示出了卓越的效率。具体而言，在10M令牌时，oba在注意力计算时间中的加速比率降低了16倍。 Thetop图中的插图图，重点是较短的序列（32K至512K），表明，即使两种方法都执行了相当的ATSMALLER尺度，但随着序列的增长，MOBA的计算优势越来越明显，随着序列的增长，突出显示了处理非常长的序列时特定的强度。

总体而言，MOBA的高效率可以归因于两个关键的创新：（1）块稀疏的注意力机械和（2）（2）相结合了Experts（MOE）和FlashEntention的优化实现，如第2.3节所述。这些技术有效地解决了全面注意力的二次复杂性限制，从而将计算复杂性降低到更经济的次级量表。

4 相关工作

有效关注的发展（Tay，Dehghani等，2020）机制一直是自然语言处理领域的关键领域，尤其是随着大语言模型（LLMS）的兴起。随着对处理较长序列和降低计算成本的需求的增长，Efficeint注意技术已成为一种有希望的解决方案，可在维持模型实力的同时降低自我注意机制的二次复杂性。

静态稀疏模式：巨大的努力，例如稀疏变压器（Child等，2019），Star-Transformer（Q. Guoetal。2019），Blockbert（Qiu等人，2019年），Longformer（Beltagy等人2020），GMAT （Gupta等，2020）等（Ainslie，Ontanon等，2020），Bigbird（Zaheer等，2020）， Longt5（M。Guo等，2021）和Longnet（J。Ding等，2023）已致力于LLMS中静态注意模式的设计。他们对静态注意力模式的选择Canencompass在关注和固定的注意力，窗户注意力，全球令牌注意力，随机关注，膨胀注意力，阻止稀疏注意力或它们的任何组合。在多模型模型的领域中，还开发了静态的稀疏注意力机制，例如2D图像和空间 - 周期性倾向（Z. Zheng等，2024）的轴向注意力（Ho等，2019），用于3D视频。

动态稀疏模式：与静态模式不同，动态稀疏注意技术自适应地确定要参加的令牌。改革者（Kitaev等，2020）和路由变压器（Roy等，2021）分别对雇用敏感性的哈希（LSH）和K-均值集群代币，并参与群集，而不是完整的环境。 Al。 K-Nearest-neighbor（KNN）算法。 Colt5（Ainslie，Lei等，2023）设计了一个路由模型来选择最重要的查询和键。稀疏的sindhorn注意力（Tay，Bahri等，2020）从输入序列中学习topermute块，从而使动态块稀疏注意计算。

无训练的稀疏注意力：除了先前讨论过的研究方法培训稀疏的方法模型外，还旨在融合稀疏注意机制，以提高模型推断的两个主要阶段的效率 - 预填充阶段或解码阶段，或两个阶段。在爆炸优化阶段，可以将完整的提示用于注意分析，从而允许对更复杂的稀疏注意力模式进行探索。例如，MOA（T. Fu等，2024），终结（H. Jiang等，2024）和seerattention（Y. Gao等，2024）已经研究了稀疏的注意力构型，例如A形，垂直斜线，和动态块稀疏性。在解码优化的背景下，大量的工作是专门的，并修剪了KV-CACHE，以在文本生成的质量和速度之间达到平衡。该领域的著名富裕物包括H2O（Z. Zhang等，2024），Streamingllm（G。Xiao等，2023），Tova（Oren等，2024），（Ge等人2023）和Quest（Tang等。 2024）。尤其可以将任务视为具有较小尺寸的MOBA和专门的块表示功能，可将Min和Max Pool结合起来。另一项工作

closely related to MoBA is Longheads (Y. Lu et al. 2024) which can be viewed as MoBA with a top-1 gating network, meaning that each query selects the most relevant KV blocks for attention.

除了传统的关注体系结构：另一条研究线研究了新型模型体系结构，从而脱离常规注意机制。随着体系结构的变化，这些方法需要从划痕进行培训模型，并且无法重复使用预训练的基于变压器的模型。该领域的研究探索了受卷积神经网络（CNNS），经常性神经网络（RNN），状态空间模型（SSM）或线性注意的启发的探索结构（Katharopoulos等，2020）。 2020），RWKV Peng，Alcaide等。 2023年，Mamba（Gu等，2023），Retnet（Sun等，2023），等。

总而言之，有效的注意力技术的景观是多种多样的，涵盖了稀疏模式，这些模式从静态到动态，优化目标，从训练到推理，以及从传统注意力机制延伸的架构创新替代方案。每种方法都呈现独特的优势和权衡，而技术的选择通常取决于应用程序的特定要求，例如最大序列长度，计算资源以及效率和性能之间所需的平衡。随着该领域的研究不断发展，预计这些方法将在使LLMS能够解决效率和可扩展性的同时解决复杂的任务时起着至关重要的作用。

5 结论

在本文中，我们介绍了块引起关注（MOBA）的混合物，这是一种受专家混合（MOE）的启发的新型注意力结构，旨在提高长语言模型（LLMS）的效率和可扩展性（LLMS）的效率和可扩展性。任务。 MOBA通过将上下文将上下文划分为块并采用动态的门控机制来选择性地将QueryTokens路由到最相关的KV块来解决与传统注意力机械态相关的计算挑战。这种方法不仅降低了计算复杂性，还可以保持模型性能。此外，它允许在全面和稀疏注意力之间进行无缝过渡。通过广泛的发挥作用，我们证明了MOBA可以达到与全部注意力相当的性能，同时显着提高了计算效率。我们的结果表明，MOBA可以有效地扩展到较长的上下文，维持各种基准的低损失和高性能。此外，MOBA的灵活性使其可以与现有模型集成在一起，而无需大量的培训成本，这使其成为增强LLMS的长期培训能力的实际持续培训解决方案。总之，MOBA代表了有效关注的重大进步，在绩效和效率之间提供了平衡的方法。未来的工作可能会探讨Moba的块选择策略的进一步优化，研究其对其他模式的应用，并研究其在复杂推理任务中改善综合性的潜力。

References

Ainslie，Joshua，Tao Lei等。 “ COLT5：具有条件计算的更快的远程变压器”。在：Arxivpreprint arxiv：2303.09752（2023）。 “ ETC：在变压器中编码长长的结构化输入”。在：arxivpreprint arxiv：2004.08483（2020）。 “为什么有效的LLM的有效上下文长度不足？”在：ARXIV预印型ARXIV：2410.18745（2024）。引入100K上下文窗口。https://www.anththropic.com/news/100k-context-windows.2023.beltagy，iz，Matthew E Peters和Arman Cohan。 “ longformer：长期文档变压器”。在：Arxiv Preprintarxiv：2004.05150（2020）。Bertsch，Amanda等。 “无形者：无限长度输入的远程变压器”。在：神经信息处理系统的进展36（2024）.Bike，Aviv等。 “变压器到SSM：将二次知识蒸馏到次级模型中”。在：进步的中学信息处理系统37（2025），第31788–31812页。Chen，Shouyuan等。 “通过位置插值扩展大语模型的上下文窗口”。在：Arxivpreprint Arxiv：2401.06066（2023）。Child，Rewon等。 “用稀疏变压器生成长序列”。在：ARXIV预印Arxiv：1904.10509（2019）。Choromanski，Krzysztof等。 “与表演者重新考虑注意力”。在：Arxiv预印ARXIV：2009.14794（2020）.dai，Damai等。 “ DeepSeekmoe：致力于终极专家专业专业，以融合了专家的语言模型”。在：ARXIV预印ARXIV：2401.06066（2024）。

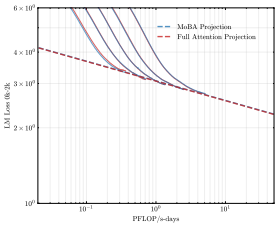
Dao，Tri，Dan Fu等。“ Flashateention：具有IO-IS-IS-IS-IS-ISISES的快速和记忆力的精确关注”。在：Advancesin神经信息处理系统35（2022），第16344–16359.Dao，Tri和Albert Gu。“变压器是SSM：通过结构性恒定空间双重性的广义模型和有效算法”。在：Arxiv预印型ARXIV：2405.21060（2024）。Ding，Jiayu等。“ Longnet：将变压器扩展到1,000,000,000代币”。在：Arxiv预印型ARXIV：2307.02486（2023）。Fedus，William，Barret Zoph和Noam Shazeer。“开关变压器：使用简单和有效的稀疏性缩放到万亿个参数模型”。在：机器学习研究杂志23.120（2022），第1-39.FU，Tianyu等。“ MOA：自动大型语言模型压缩的稀疏注意力的混合物”。在：Arxivpreprint Arxiv：2406.14909（2024）。Gao，Yizhao等。“宣传：在LLM中学习固有的稀疏关注”。在：Arxiv Preprintarxiv：2410.13276（2024）。GE，Suyu等。“ Model告诉您要丢弃什么：LLMS的自适应KV缓存压缩”。在：arxiv preprintarxiv：2310.01801（2023）.gu，Albert和Tri Dao。“ Mamba：具有选择性状态空间的线性时间序列建模”。在：Arxiv preprintarxiv：2312.00752（2023）。Guan，Melody Y等。“辩护：推理可以使语言模型更安全”。在：Arxiv Preprintarxiv：2412.16339（2024）。Guo，Daya等。“ DeepSeek-R1：通过增强学习激励LLM中的推理能力”。在：Arxivpreprint arxiv：2501.12948（2025）。Guo，Mandy等。“ longt5：长序列的有效文本到文本变压器”。在：Arxiv preprintarxiv：2112.07916（2021）。Guo，Qipeng等。“星形转换器”。在：Arxiv Preprint Arxiv：1902.09113（2019）。Gupta，Ankit和Jonathan Berant。“ GMAT：变压器的全球内存增强”。在：arxiv preprintarxiv：2006.03274（2020）。“多维变压器中的轴向注意力”。在：ARXIV预印ARXIV：1912.12180（2019）.Hoffmann，Jordan等。“培训计算最佳的大语言模型”。在：ARXIV预印型ARXIV：2203.15556（2022）.Jiang，Huiqiang等。“临界1.0：通过动态稀疏注意加速长篇文化LLM的预填充”。“变压器是RNN：具有线性注意的快速自回旋变压器”。在：机器学习国际会议上。 PMLR。 2020年，第5156–5165页。Kitaev，Nikita，Oukasz Kaiser和Anselm Levskaya。“改革者：高效的变压器”。在：Arxiv Preprintarxiv：2001.04451（2020）。Lepikhin，Dmitry等。“ GSHARD：使用条件计算和自动碎片的巨型模型”。在：Arxiv预印ARXIV：2006.16668（2020）.li，Aonian等。“ Minimax-01：具有闪电注意力的扩展基础模型”。在：arxiv preprintarxiv：2501.08313（2025）.liu，di等。“检索：通过矢量检索加速长篇小说LLM推断”。在：arxiv preprintarxiv：2409.10516（2024）.liu，hao和Pieter Abbeel。“大型上下文模型的块平行变压器”。在：arxiv preprintarxiv：2305.19370（2023）.lu，yi等。“远程：多头关注是秘密的上下文处理器”。在：Arxiv Preprintarxiv：2402.10685（2024）。Mercat，Jean等。“线性化大语言模型”。在：Arxiv预印型ARXIV：2405.06640（2024）。Milakov，Maxim和Natalia Gimelshein。“ SoftMax的在线归一化计算”。在：arxiv preprintarxiv：1805.02867（2018）.moonshotai。 Kimi聊天。 https://kimi.moonshot.cn/。 2023.oren，Matanel等。“变压器是多状态RNN”。在：ARXIV预印型ARXIV：2401.06104（2024）.Peng，Bo，Eric Alcaide等。“ RWKV：重塑变压器时代的RNN”。在：Arxiv预印型ARXIV：2305.13048（2023）.Peng，Bo，Daniel Goldstein等。“ Eagle and Finch：具有矩阵值态和动态复发的RWKV”。在：ARXIV预印ARXIV：2404.05892（2024）.Poli，Michael等。“鬣狗的层次结构：迈向更大的卷积语言模型”。在：国际会议机器学习。 PMLR。 2023年，第28043–28078页，Qiu，Jiezhong等。“对长期文档理解的自我注意力”。在：ARXIV预印Arxiv：1911.02972（2019）.Rajbhandari，Samyam等。“ DeepSpeed-Moe：推进专家的混合物推理和训练为下一代AI量表提供动力”。在：机器学习国际会议上。 PMLR。 2022年，第18332– 18346年。

里德，马切尔等。“双子座1.5：在数百万个上下文中解锁多模式理解”。在：Arxivpreprint Arxiv：2403.05530（2024）。Roy，Aurko等。“通过路由变压器进行有效的基于内容的稀疏注意力”。在：计算语言学协会的交易9（2021），第53-68页。Shazeer，Noam等。“令人毛骨悚然的大神经网络：稀疏门控的专家层”。在：Arxivpreprint Arxiv：1701.06538（2017）。Shoeybi，Mohammad等。“ Megatron-LM：使用模型 - 模型训练数十亿个参数语言模型”。在：ARXIV预印ARXIV：1909.08053（2019）。Sun，Yutao等。“保留网络：大型语言模型的变压器的继任者”。在：arxiv preprintarxiv：2307.08621（2023）。“任务：查询意识到的稀疏性对于有效的长篇小说LLM推理”。在：arxiv preprintarxiv：2406.10774（2024）.tay，yi，dara Bahri等。“稀疏的sindhorn注意力”。在：机器学习国际会议上。 PMLR.2020，第9438–9447页。“有效的变压器：一项调查。 arxiv”。在：ARXIV预印ARXIV：2009.06732（2020）。Team，Kimi等。“ Kimi K1。 5：使用LLMS缩放加强学习”。在：ARXIV预印型ARXIV：2501.12599（2025）。Wang，Junxiong等。“骆驼中的曼巴：蒸馏和加速混合模型”。在：Arxiv Preprintarxiv：2408.15237（2024）。Wang，Sinong等。“线形：具有线性复杂性的自我注意力”。在：Arxiv预印ARXIV：2006.04768（2020）.Waswani，A等。“关注就是您需要的”。在：nips。 2017年。Watson，Jake F等。“人类海马CA3使用特定的功能连通性规则来有效地关联”。在：Cell 188.2（2025），第501–514.Wu，Yuhuai等。“记忆变压器”。在：Arxiv预印ARXIV：2203.08913（2022）.xiao，Guangxuan等。“有效的流媒体语言模型，带有注意力降低”。在：arxiv preprintarxiv：2309.17453（2023）.xiong，Wenhan等。“基础模型的有效长篇小说缩放”。在：Arxiv预印型ARXIV：2309.16039（2023）.Yang，An等。“ qwen2。 5个技术报告”。在：ARXIV预印型ARXIV：2412.15115（2024）.Zaheer，Manzil等。“大鸟：更长序列的变压器”。在：神经信息处理系统的进步33（2020），第17283–17297页。Zhang，Michael等。“ LOLCATS：大型语言模型的低排名线性化”。在：arxiv preprintarxiv：2410.10254（2024）.zhang，Xuan等。“ Simlayerkv：层级kV缓存降低的简单框架”。在：arxiv preprintarxiv：2410.13846（2024）.zhang，Zhenyu等。“ H2O：重击甲骨文，以有效地推断大语言模型”。在：神经信息处理系统中的广告范围36（2024）.Zheng，Zangwei等。开放式：使所有人的有效视频制作民主化。2024年3月。URL：https：//github.com/hpcaitech/open-sora.zoph，Barret等。“ St-Moe：设计稳定且可转移的稀疏专家模型”。在：Arxiv Preprintarxiv：2202.08906（2022）。

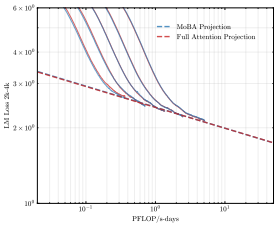
附录

A.1长上下文可伸缩性

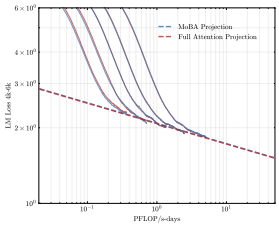
为了解决有利于简短上下文的自然数据分布的偏见，我们根据其实际位置将整体观点分为离散的细分市场。例如，跨越位置的段30k-32二额反映了与超过30k上下文长度的文档相关的损失，并且还掩盖了从30k到32k的位置。这种方法可确保在不同上下文长度上进行更平衡和代表性的评估。 INOUR探索长篇小说可伸缩性，我们进行了一个关键的发现：尾随令牌是整个上下文基线与新提出的新提出的稀疏注意架构之间大部分性能差异的说明。事实称，我们通过将重点放在远处的缩放过程中来简化了长篇文章的缩放过程落后令牌缩放。这不仅可以简化计算要求，而且还可以显着提高投资通用长篇小说方案的效率和有效性。这一发现对未来更有效，更尺寸的注意机制的发展具有重大意义。



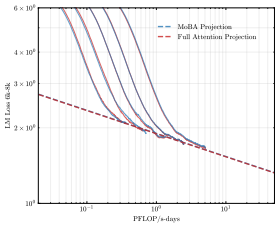
(a) Scaling law (0-2k)



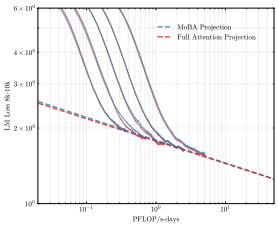
(b) Scaling law (2-4k)



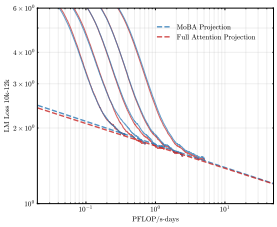
(c) Scaling law (4-6k)



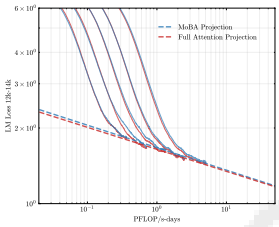
(d) Scaling law (6-8k)



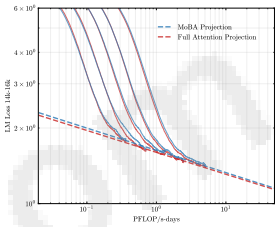
(e) Scaling law (8-10k)



(f) Scaling law (10-12k)

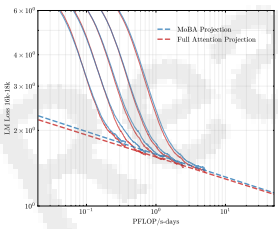


(g) Scaling law (12-14k)

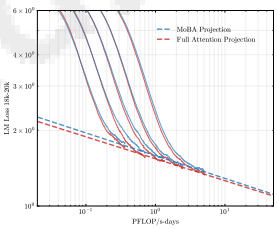


(h) Scaling law (14-16k)

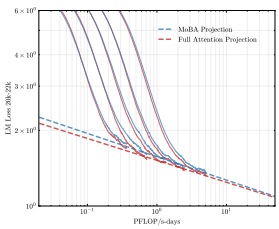
图8：位置的缩放定律0-16K



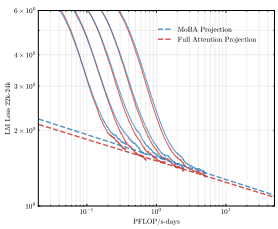
(i) Scaling law (16-18k)



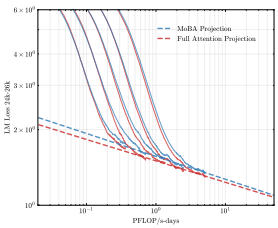
(j) Scaling law (18-20k)



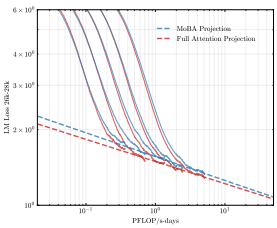
(k) Scaling law (20-22k)



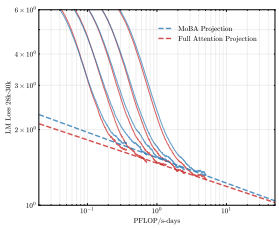
(l) Scaling law (22-24k)



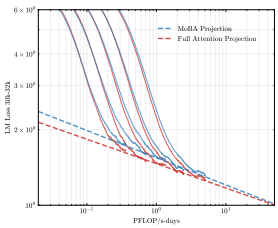
(m) Scaling law (24-26k)



(n) Scaling law (26-28k)



(o) Scaling law (28-30k)



(p) Scaling law (30-32k)

图8：位置的缩放定律16-32K

表3：具有不同位置的损失缩放

LM Loss Position Range	MoBA	Full
0K - 2K	$3.075 \times C^{-0.078}$	$3.068 \times C^{-0.078}$
2K - 4K	$2.415 \times C^{-0.084}$	$2.411 \times C^{-0.083}$
4K - 6K	$2.085 \times C^{-0.081}$	$2.077 \times C^{-0.081}$
6K - 8K	$1.899 \times C^{-0.092}$	$1.894 \times C^{-0.092}$
8K - 10K	$1.789 \times C^{-0.091}$	$1.774 \times C^{-0.089}$
10K - 12K	$1.721 \times C^{-0.092}$	$1.697 \times C^{-0.087}$
12K - 14K	$1.670 \times C^{-0.089}$	$1.645 \times C^{-0.088}$
14K - 16K	$1.630 \times C^{-0.089}$	$1.600 \times C^{-0.087}$
16K - 18K	$1.607 \times C^{-0.090}$	$1.567 \times C^{-0.087}$
18K - 20K	$1.586 \times C^{-0.091}$	$1.542 \times C^{-0.087}$
20K - 22K	$1.571 \times C^{-0.093}$	$1.519 \times C^{-0.086}$
22K - 24K	$1.566 \times C^{-0.089}$	$1.513 \times C^{-0.085}$
24K - 26K	$1.565 \times C^{-0.091}$	$1.502 \times C^{-0.085}$
26K - 28K	$1.562 \times C^{-0.095}$	$1.493 \times C^{-0.088}$
28K - 30K	$1.547 \times C^{-0.097}$	$1.471 \times C^{-0.091}$
30K - 32K	$1.546 \times C^{-0.108}$	$1.464 \times C^{-0.097}$