

MOBA

LLM

Enzhe Lu<sup>1</sup>   Zhejun Jiang<sup>1</sup>   Jingyuan Liu<sup>1</sup>   Yulun Du<sup>1</sup>   Tao Jiang<sup>1</sup>   Chao Hong<sup>1</sup>  
Shaowei Liu<sup>1</sup>   Weiran He<sup>1</sup>   Enming Yuan<sup>1</sup>   Yuzhi Wang<sup>1</sup>   Zhiqi Huang<sup>1</sup>   Huan Yuan<sup>1</sup>  
Suting Xu<sup>1</sup>   Xinran Xu<sup>1</sup>   Guokun Lai<sup>1</sup>   Yanru Chen<sup>1</sup>   Huabin Zheng<sup>1</sup>   Junjie Yan<sup>1</sup>  
Jianlin Su<sup>1</sup>   Yuxin Wu<sup>1</sup>   Neo Y. Zhang<sup>1</sup>   Zhilin Yang<sup>1</sup>  
Xinyu Zhou<sup>1,z</sup>   Mingxing Zhang<sup>2,y</sup>   Jiezhong Qiu<sup>3,z,y</sup>

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

ABSTRACT

LLM

AGI

com-plex

"

"

MOBA

MOE  
su-prorior

MOBA

Kimi

LLMS

<https://github.com/moonshotai/moba>

1

AGI

LLMS

AGI

Kimi   Moonshotai 2023   Claude   Anthropic 2023   Gemini   Reid   Reid  
etal   2024   Team   2025   DeepSeek-R   D. Guo   2025   Openai O /O   Guan   2024  
Kimi K .   COT

Waswani   2017

LLMS

H. Jiang   2024   Watson   2025

Wherespase

zhang\_mingxi ng@mail . tsi nghua. edu. cn

<sup>y,z</sup>Co-corresponding authors. Xinyu Zhou (zhouxi nyu@moonshot. cn), Jiezhong Qiu (j i ezhongqi u@outl ook. com)

Beltagy 2020 G. Xiao 2023  
 2024 Quest Tang 2024 H. Jiang  
 Presprence t. Di Liu 2024 Tokens of Potkens of  
 LLM  
 2023; Peng Peng Goldstein 2024 Mamba Dao Gu 2024 RWKV Peng Alcaide  
 Retnet Sun et et and Sun et et nechte AI  
 2024; Bick etal 2025; M. Zhang etal 2024; Mercat 2024; J. Wang  
 2025 A. Li

" "

Shazeer 2017 MOBA  
 2022 MOBA FFN Lepikhin TransformerModel MOE  
 2020 Fedus 2022 Zoph MOE  
 LLM  
 MOBA  
 mostretlevant

MOBA  
 MOBA  
 LLM

2

" MOBA "  
 CA-PIS MOE  
 FFN

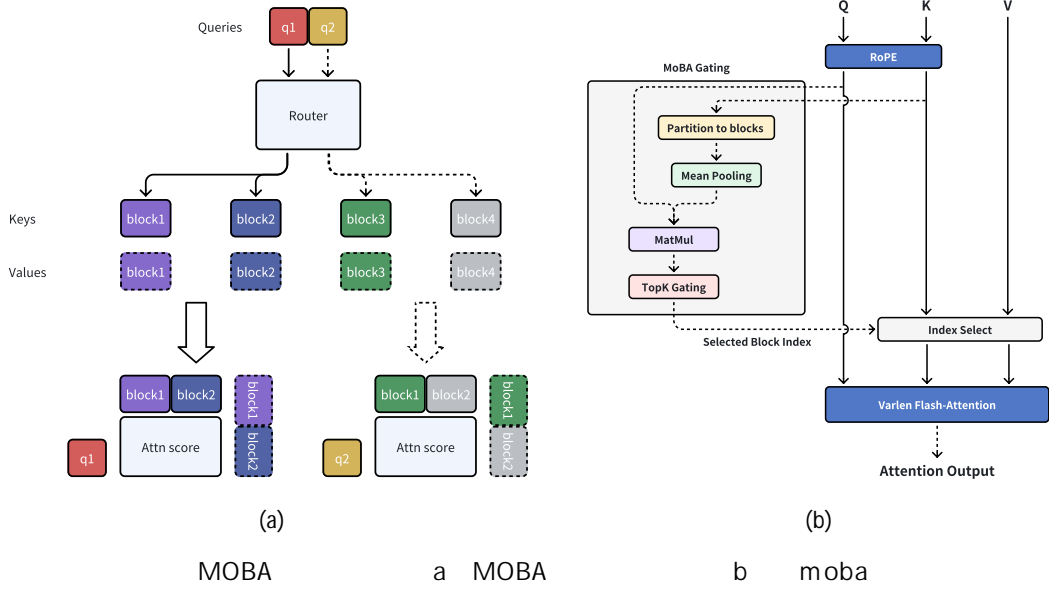
2.1

$$\text{token}_q \cdot R \times d$$

$$n \quad k \quad v \quad Rn \times d$$

$$\text{attn}_q \cdot k \cdot v = \text{softmax}(qk^T) \cdot v \quad 1 \quad (1)$$

D



## 2.2 MoBA Architecture

MOBA

$$\text{moba} \quad q \quad k \quad v = \text{softmax} \quad qk \quad [i] \quad v \quad [i] \quad 2 \quad > \quad V[I]; \quad (2)$$

$i \quad [n]$

MOBA

$n$

$n$

$n$

$$b = nn$$

$$\text{and } i = [ \quad i - 1 \quad \times b + 1 \quad i \times b ] \quad 3$$

MOE TOP-K

$$I = \begin{bmatrix} I_1 \\ I_2 \\ \vdots \\ I_n \end{bmatrix} \quad (4)$$

$Q \quad i\text{-th}$

$4 \quad G \quad i$

- top-k

MOBAGate

$I\text{-th Block } G \quad I$

$S \quad I$

$$g_i = \begin{cases} 1 & s_i \geq \text{Topk} (fs_{jj} \quad 2 \quad [n]g; k) ; \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

$\text{topk} \quad \cdot \quad k$   
 $s \quad Q \quad k \quad [i]$

$k$

$$s \quad i = q$$

$$k \quad [i \quad i]$$

6

1A

MOBA

1A

KV

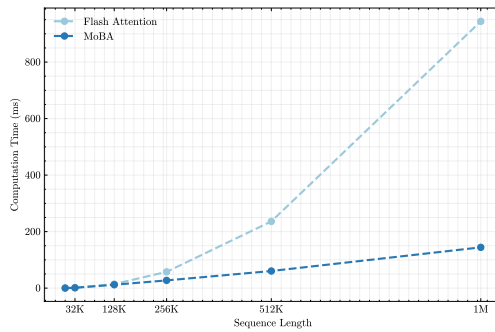


1 MOBA

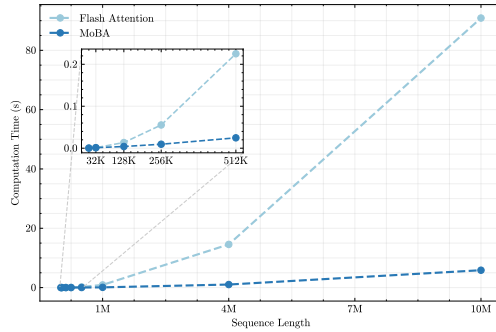
```

q k v Rn×h×d; MOBA B B TOP-K D
n = n/b blocks. // kv blocks {ki i} =
splitblocks k k v b ki i l Rbb×h×d i [n] 3 //
4 k = k b Rn×h×d s=qk Rn×h×n //
7 M = n n 8 g=topk s+m k 9 //
10 qs ks s= attn q k 11 qm km m=index select
selectoba attn q k g 12 // 13 os= ash fastion varlen qs
ks s Causal = true 14 om = ash faster varlen qm km km km km
M = false 15 // SoftMax o = SoftMax OS OM 17
o

```



(a)



(b)

2 MOBA

1M

a M

MOBA

MOBA Flash

95.31

64 MOBA Blockswith Variancation Block Block

8K - M

b 8K - M

TOP-K = 3

•

• SoftMax

kv

Flashingention 1 MOBA

1B KV kvblocks

1-2 3-7

MOE G

6

6 TOP-K

SPARSECORY-KV-BLOCK QueryTokens

KV 8 Line -

11 14

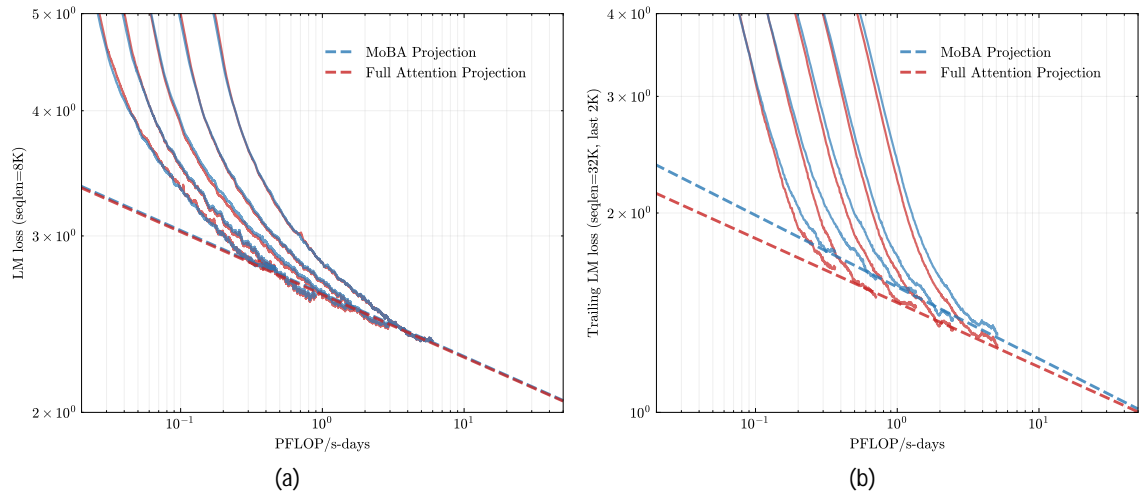
10 13

16

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



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

(c)

3  
c

MOBA                      a                      LM                      seqen = k ;                      b                      LM                      seqen = k                      1k

3

3.1

MOBA

W.R.T. LM                      MOBA                      Chinchilla Scalinglaw                      Ho mann                      MOBA 2022

MOBA                      8K                      Mobamodels

512                      3                      MOBA

1-512× 38192 = 81.25                      3

3A                      MOBA

1E -3

75                      MOBA

3                      top-k = 3                      2

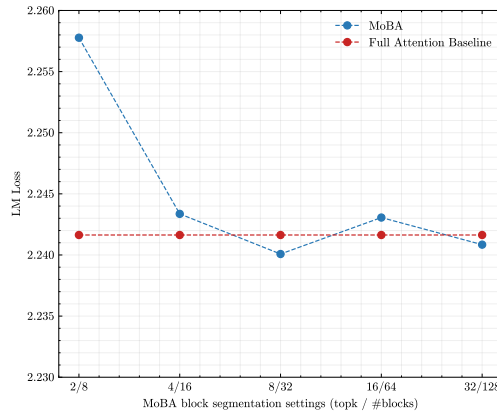
LM MOBA An 2024  
LM A.

8K 32K MOBA  
1-512x 332768 = 95.31 LM Mobaexhib 3B

MOBA

1.5B MOBA 32K  
32k Contextln to 8 16 32 64 128 TOP-K 2 4 8 16 32  
75 4 MOBA  
8 2 1E-

MOBA MoE



4 LM v.s MOBA

3.2 MOBA

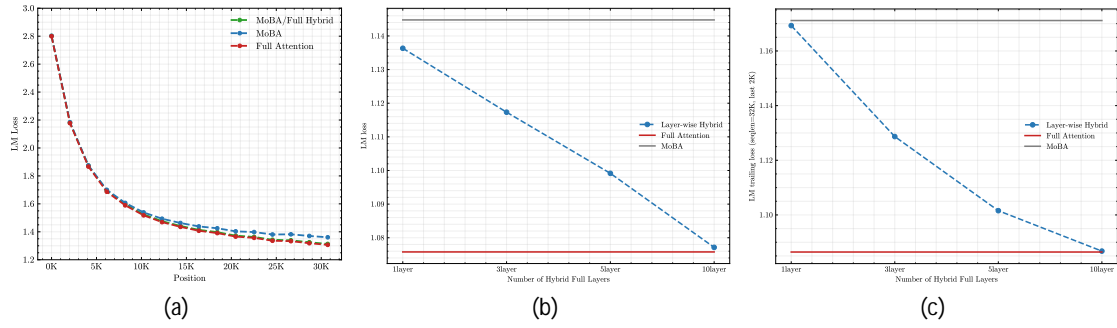
2 MOBA We rst MOBA / SFT /

MOBA / 30B 1.5B  
32K MOBA 2048 Top-K Setto 3

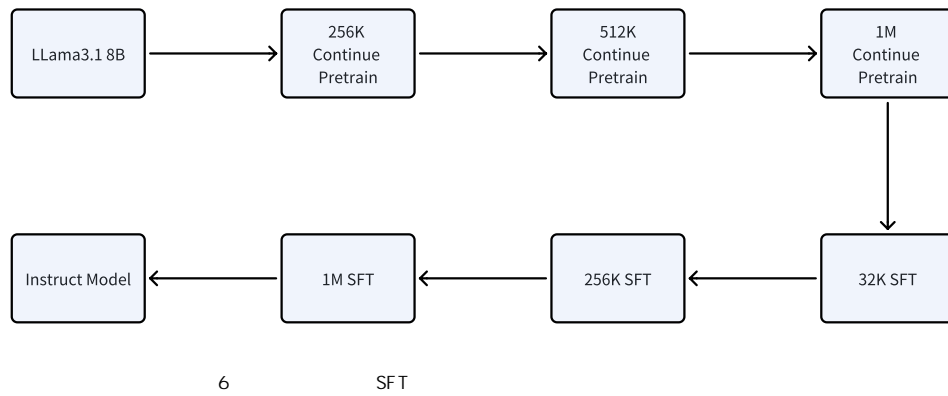
• MOBA/FULL HYBRID  
90

10 • MOBA MOBA

LM LM LM LM LM  
2024



5 MOBA a MOBA LM MOBA/Full Hybrid b sft 1m W.R.T  
c sft tailding lmloss seq len = k 2k w.r.t these Hybrid



LM 5A MOBA  
MOBA/FULL HYBRID  
MOBA/Full Hybrid TrainingRecipe  
MOBA Moba  
MOBA - thelayer MOBA  
SFT SFT 5B  
SFT MOBA  
eContext  
MOBA 5B 5C Moba  
SFT

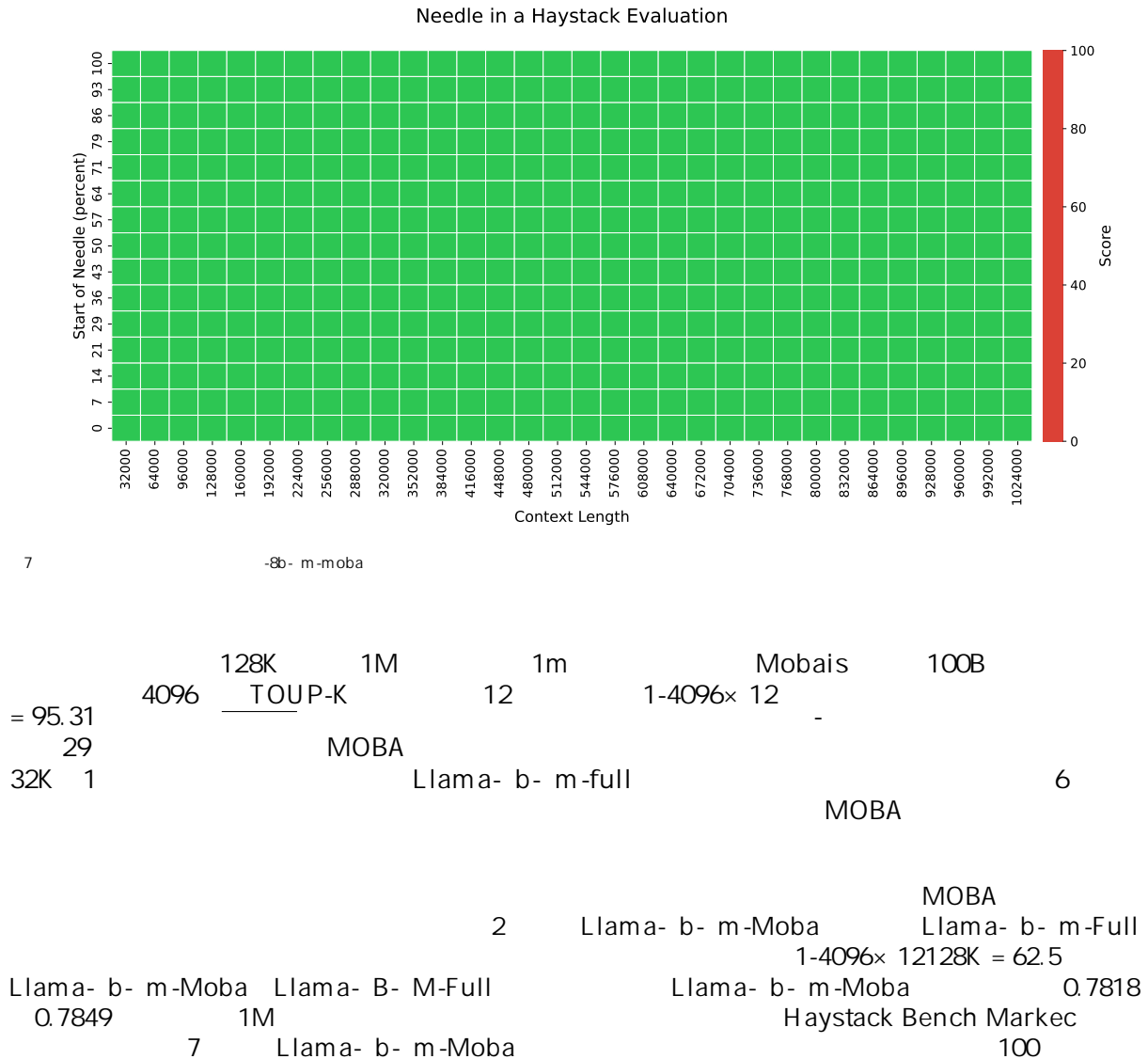
3.3

MOBA  
Llama . B  
Llama- B- M-MOBA 128K  
TO K K 1M  
S. Chen 2023 256K



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

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



3.4

MOBA  
3.3  
theattention  
FFN  
2A  
1M  
1000  
MOBA  
K  
Shoeybi etal  
2019  
GPU  
MOBA  
10M  
oba  
32K  
512K  
MOBA  
MobAdjext  
16  
Thetop  
ATSMALLER  
MOBA  
6.5  
10m  
2B  
MOBA  
Experts  
MOE  
FlashEntention  
1  
2.3  
2  
2

4

Tay Dehghani 2020  
LLMS  
E ceint  
2019 Blockbert Qiu 2019 Child 2019 Star-Transformer Q. Guoetal  
2020 Ainslie Ontanon 2020 Longformer Beltagy 2020 GMAT Gupta  
2021 Longnet J Ding 2023 Bigbird Zaheer 2020 Longt M Guo  
Canencompass  
- Z. Zheng 2024 Ho 2019 3D  
2020 Roy 2021 LSH K-  
Colt Ainslie Lei 2023 AI K-Nearest-neighbor KNN  
Tay Bahri 2020 topermute  
-  
MOA T. Fu 2024 H. Jiang 2024 seerattention Y. Gao 2024  
A  
KV-CACHE  
H O Z. Zhang 2024 Streamingllm G Xiao 2023 Tova Oren  
2024 Ge 2023 Quest Tang 2024  
MOBA Min Max Pool

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.

SSM				CNNs				RNN	RWKV
Peng	Alcaide	2023	Mamba	Gu	Katharopoulos	2020	2020		
					2023	Retnet	Sun	2023	

LLMS

5

[illegible]

## References

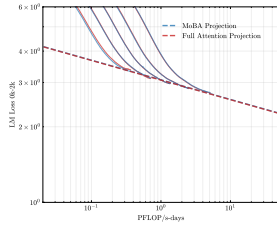
Inslie Joshua Tao Lei	"	COLT	"	
Arxivpreprintarxiv	2303.09752	2023	"	ETC
"	arxivpreprintarxiv	2004.08483	2020	"
"	ARXIV	ARXIV	2410.18745	2024
https://www.anthropic.com/news/	k-context-windows.	.beltagy	iz	Matthew
E Peters Arman Cohan	"	longformer	"	Arxiv
Preprintarxiv	2004.05150	2020	Bertsch Amanda	"
"		36	2024	.Bike Aviv
"				37 2025
Chen Shouyuan	"			31788-31812
ArxivpreprintArxiv	2401.06066	2023	Child Rewon	"
"	ARXIV	Arxiv	1904.10509	2019
"	Arxiv	ARXIV	.	.dai Damai
DeepSeekmoe			"	ARXIV
ARXIV	2401.06066	2024		

Dao Tri Dan Fu " Flashateention IO-IS-IS-IS-IS-ISIS  
 " Advancesin 35 2022 16344-16359.Dao Tri Albert  
 Gu " SSM " Arxiv  
 ARXIV 2405.21060 2024 Ding Jiayu " Longnet 1,000,000,000  
 " Arxiv ARXIV 2307.02486 2023 Fedus William Barret Zoph  
 Noam Shazeer " "  
 23.120 2022 1-39.FU Tianyu " MOA  
 " Arxivpreprint Arxiv 2406.14909 2024 Gao Yizhao  
 " LLM " Arxiv Preprint arxiv 2410.13276 2024  
 GE Suyu " Model LLMS KV " arxiv  
 preprint arxiv . gu Albert Tri Dao " Mamba  
 " Arxiv preprint arxiv 2312.00752 2023 Guan Melody  
 Y " " Arxiv Preprint arxiv  
 2412.16339 2024 Guo Daya " DeepSeek-R LLM  
 " Arxivpreprint arxiv 2501.12948 2025 Guo Mandy " longt  
 " Arxiv preprint arxiv 2112.07916 2021 Guo Qipeng  
 " " Arxiv Preprint Arxiv 1902.09113 2019 Gupta Ankit  
 Jonathan Berant " GMAT " arxiv preprint arxiv  
 2006.03274 2020 " " ARXIV ARXIV  
 . Ho mann Jordan " " ARXIV  
 ARXIV . Jiang Huiqiang " 1.0  
 LLM " " RNN  
 PMLR 2020 5156-5165 Kitaev Nikita Oukasz Kaiser  
 Anselm Levskaya " " Arxiv Preprint arxiv  
 2001.04451 2020 Lepikhin Dmitry " GSHARD  
 " Arxiv ARXIV . li Aonian " Minimax-  
 " arxiv preprint arxiv . liu di  
 " LLM " arxiv preprint arxiv  
 . liu hao Pieter Abbeel " "  
 arxiv preprint arxiv . lu yi " "  
 " Arxiv Preprint arxiv 2402.10685 2024 Mercat Jean " "  
 " Arxiv ARXIV 2405.06640 2024 Milakov Maxim Natalia  
 Glimelshein " SoftMax arxiv preprint arxiv  
 . moonshotai Kimi https://kimi.moonshot.cn/ .oren  
 Matanel " RNN" ARXIV ARXIV  
 . Peng Bo Eric Alcaide " RWKV RNN"  
 Arxiv ARXIV . Peng Bo Daniel Goldstein " Eagle  
 and Finch RWKV" ARXIV ARXIV  
 . Poli Michael " "  
 PMLR 2023 28043-28078 Qiu Jiezhong "  
 " ARXIV Arxiv . Rajbhandari  
 Samyam " DeepSpeed-Moe AI  
 " PMLR 2022 18332- 18346

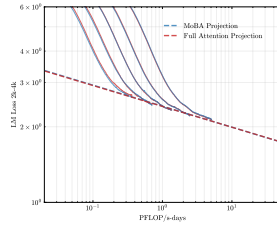
Arxivpreprint	Arxiv	2403.05530	2024	Roy Aurko	"	"	53-68
Shazeer Noam	"	"	"	"	"	"	"
Arxivpreprint	Arxiv	1701.06538	2017	Shoeybi Mohammad	"	Megatron-LM	"
1909.08053	2019	Sun Yutao	"	"	"	"	"
arxiv preprint	arxiv	2307.08621	2023	"	"	"	"
LLM	"	arxiv preprint	arxiv	"	"	"	"
Bahri	"	sindhorn	"	"	"	"	"
9447	"	"	"	"	"	"	"
2009.06732	2020	Team Kimi	"	Kimi K 5	LLMS	"	"
ARXIV	ARXIV	2501.12599	2025	Wang Junxiong	"	"	"
"	"	Arxiv Preprint	arxiv	2408.15237	2024	Wang	"
Sinong	"	"	"	"	"	"	"
"	"	"	"	"	"	"	"
Watson Jake F	"	CA	"	"	"	"	"
188.2	2025	501-514.Wu Yuhuai	"	"	"	"	"
"	"	xiao Guangxuan	"	"	"	"	"
"	"	arxiv preprint	arxiv	"	"	"	"
"	"	Arxiv	ARXIV	"	"	"	"
qwen 5	"	ARXIV	ARXIV	"	"	"	"
Manzil	"	"	"	"	"	"	"
17283-17297	Zhang Michael	"	LOLCATS	"	"	"	"
arxiv preprint	arxiv	"	"	"	"	"	"
kV	"	arxiv preprint	arxiv	"	"	"	"
Zhenyu	"	H O	"	"	"	"	"
2024 3	URL	https	//github.com/hpcaitech/open-sora.zoph	Barret	"	St-	"
Moe	"	"	"	"	"	"	"
2202.08906	2022	"	"	"	"	"	"

A.

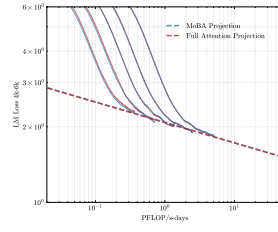
INOUR 30k 32k 30k - 30k



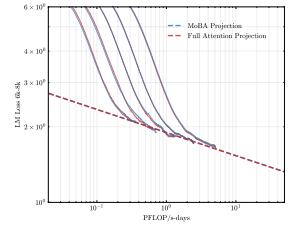
(a) Scaling law (0-2k)



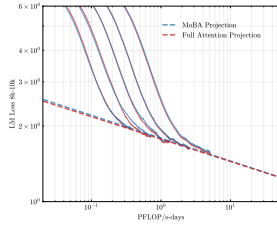
(b) Scaling law (2-4k)



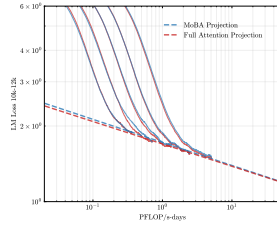
(c) Scaling law (4-6k)



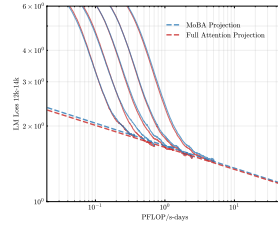
(d) Scaling law (6-8k)



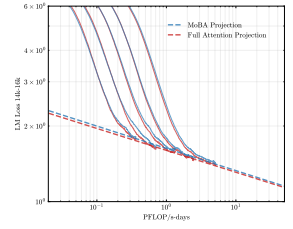
(e) Scaling law (8-10k)



(f) Scaling law (10-12k)



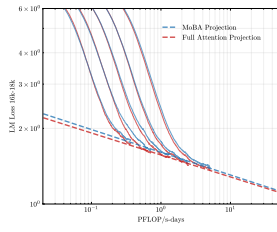
(g) Scaling law (12-14k)



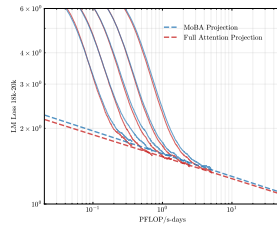
(h) Scaling law (14-16k)

8

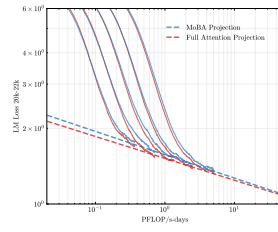
O-16K



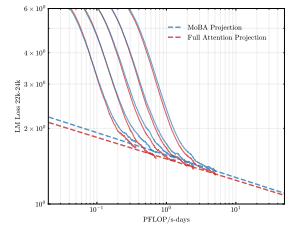
(i) Scaling law (16-18k)



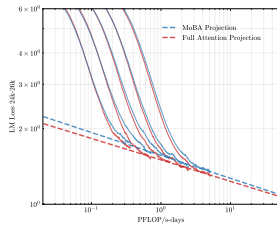
(j) Scaling law (18-20k)



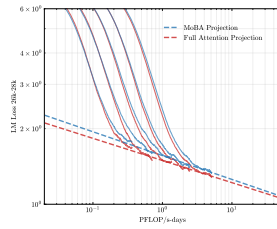
(k) Scaling law (20-22k)



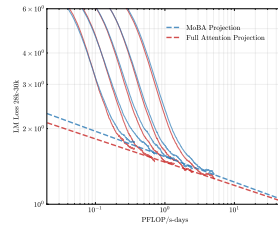
(l) Scaling law (22-24k)



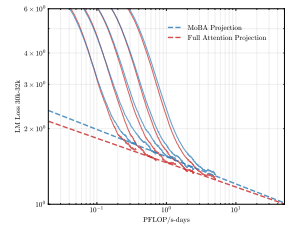
(m) Scaling law (24-26k)



(n) Scaling law (26-28k)



(o) Scaling law (28-30k)



(p) Scaling law (30-32k)

8

16-32K

3

LM Loss Position Range	MoBA		Full	
0K - 2K	3.075	<i>C</i> 0:078	3.068	<i>C</i> 0:078
2K - 4K	2.415	<i>C</i> 0:084	2.411	<i>C</i> 0:083
4K - 6K	2.085	<i>C</i> 0:081	2.077	<i>C</i> 0:081
6K - 8K	1.899	<i>C</i> 0:092	1.894	<i>C</i> 0:092
8K - 10K	1.789	<i>C</i> 0:091	1.774	<i>C</i> 0:089
10K - 12K	1.721	<i>C</i> 0:092	1.697	<i>C</i> 0:087
12K - 14K	1.670	<i>C</i> 0:089	1.645	<i>C</i> 0:088
14K - 16K	1.630	<i>C</i> 0:089	1.600	<i>C</i> 0:087
16K - 18K	1.607	<i>C</i> 0:090	1.567	<i>C</i> 0:087
18K - 20K	1.586	<i>C</i> 0:091	1.542	<i>C</i> 0:087
20K - 22K	1.571	<i>C</i> 0:093	1.519	<i>C</i> 0:086
22K - 24K	1.566	<i>C</i> 0:089	1.513	<i>C</i> 0:085
24K - 26K	1.565	<i>C</i> 0:091	1.502	<i>C</i> 0:085
26K - 28K	1.562	<i>C</i> 0:095	1.493	<i>C</i> 0:088
28K - 30K	1.547	<i>C</i> 0:097	1.471	<i>C</i> 0:091
30K - 32K	1.546	<i>C</i> 0:108	1.464	<i>C</i> 0:097