MobileLLM:
Optimizing Sub-billion Parameter Language Models for On-Device Use Cases

Zechun Liu, Changsheng Zhao, Forrest Iandola, Chen Lai, Yuandong Tian, Igor Fedorov, Yunyang Xiong,
Ernie Chang, Yangyang Shi, Raghuraman Krishnamoorthi, Liangzhen Lai, Vikas Chandra

BACKGROUND

Binary Neural Networks:
Weights and activations are binarized to –1 and +1.

Forward pass:

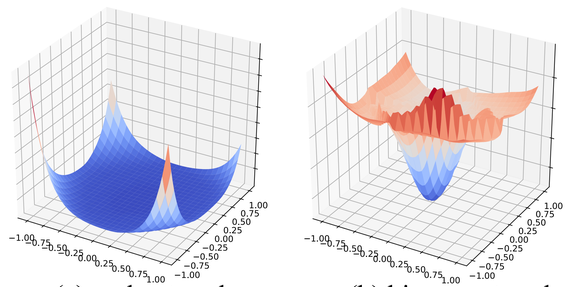
$$a_b = \text{Sign}(a_r) = \begin{cases} -1 \text{ if } & a_r < 0 \\ +1 \text{ otherwise} \end{cases}$$

Backward pass:

$$\frac{\partial Sign(a_r)}{\partial a_r} \approx \frac{\partial Clip(-1, a_r, 1)}{\partial a_r} = \begin{cases} 1 & -1 < a_r < 1 \\ 0 & \text{otherwise} \end{cases}$$

CHALLENGES OF BNNS

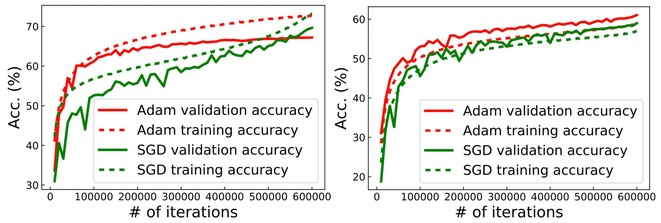
BNNs have sharper local minima compared to the real-valued networks



(a) real networks (b) binary networks

- Because BNNs restrict the weights and activations to discrete values (-1, +1), which naturally limits the representational capacity of the model and further result in disparate optimization landscapes.
- These properties differentiate BNNs from real-valued networks and impact the optimal optimizer and training strategy design.

OPTIMIZATION CURVE

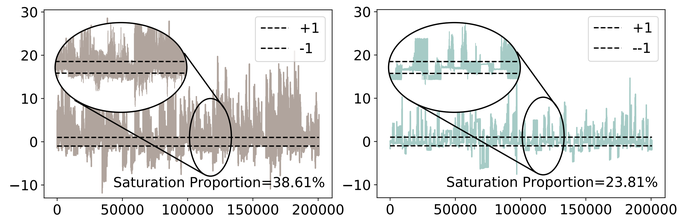


(a) real networks (b) binary networks

- On the real-valued network, SGD achieves higher accuracy with better generalization ability in the final few iterations.
- While Adam outperforms SGD for BNNs.

WHY ADAM IS BETTER FOR BNNS

Activation saturation result in gradient vanish.

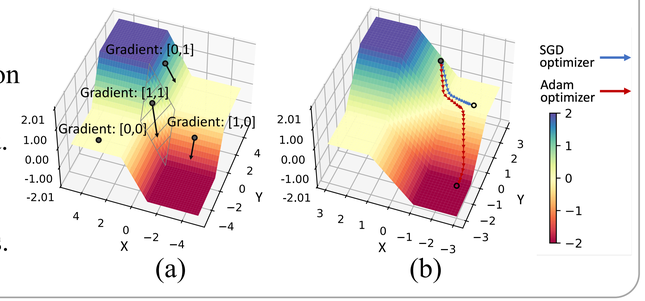


(a) SGD (b) Adam

- In the backward pass, the derivative of the sign function will encounter a zero (or vanishing) gradient problem when the activation exceeds the effective gradient range ([−1, 1]).
- SGD update: $v_t = \beta_1 v_{t-1} + g_t$
- Adam update: $m_t = \beta_2 m_{t-1} + g_t^2 \quad u_t = \frac{v_t}{\sqrt{m_t + \epsilon}}$
- Adam naturally leverages the accumulation in the second momentum to amplify the learning rate regarding the gradients with small historical values. Thus, “dead” weights from saturation are easier to be re-activated by Adam than SGD.

TWO-DIMENSIONAL OPTIMIZATION LANDSCAPE

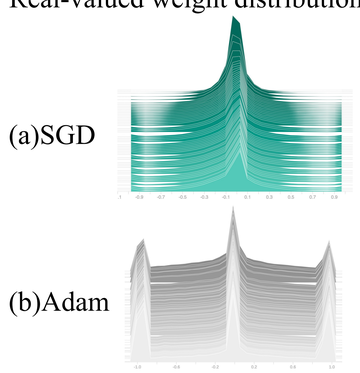
(a) The gradients only retain their values in both direction in the interval of [-1, 1], denoted as the slashed area.
(b) Adam contains higher proportion in update value when the gradient vanishes.



(a) (b)

VISUALIZATION

Real-valued weight distribution



- The magnitude of real-valued weights indicates how easy the corresponding binary weights can switch their signs (-1 or +1) to the opposite direction.
- Thus, real-valued weights can be regarded as the ***confidence*** of binary weights to be -1/ +1.

EXPERIMENTS

Dataset: ImageNet

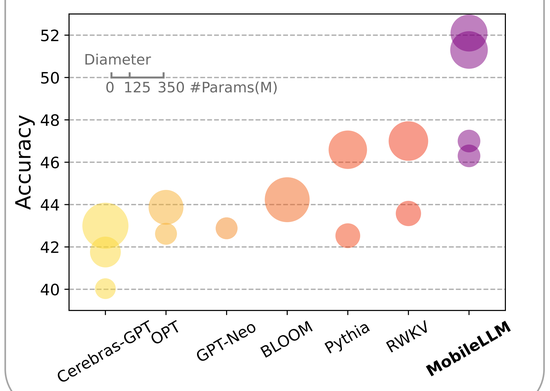
Networks	Top1 Acc %	Top5 Acc %
BNNs (Courbariaux et al., 2016)	42.2	67.1
ABC-Net (Lin et al., 2017)	42.7	67.6
DoReFa-Net (Zhou et al., 2016)	43.6	-
XNOR-ResNet-18 (Rastegari et al., 2016)	51.2	69.3
Bi-RealNet-18 (Liu et al., 2018b)	56.4	79.5
CI-BCNN-18 (Wang et al., 2019)	59.9	84.2
MoBiNet (Phan et al., 2020a)	54.4	77.5
BinarizeMobileNet (Phan et al., 2020b)	51.1	74.2
PCNN (Gu et al., 2019)	57.3	80.0
StrongBaseline (Brais Martinez, 2020)	60.9	83.0
Real-to-Binary Net (Brais Martinez, 2020)	65.4	86.2
MeliusNet29 (Bethge et al., 2020)	65.8	–
ReActNet ResNet-based (Liu et al., 2020)	65.5	86.1
ReActNet-A (Liu et al., 2020)	69.4	88.6
StrongBaseline + Our training strategy	63.2	84.0
ReActNet-A + Our training strategy	**70.5**	**89.1**

With the same architecture our training strategy bring 1.1% improvement over ReActNet.
Code is available ➞

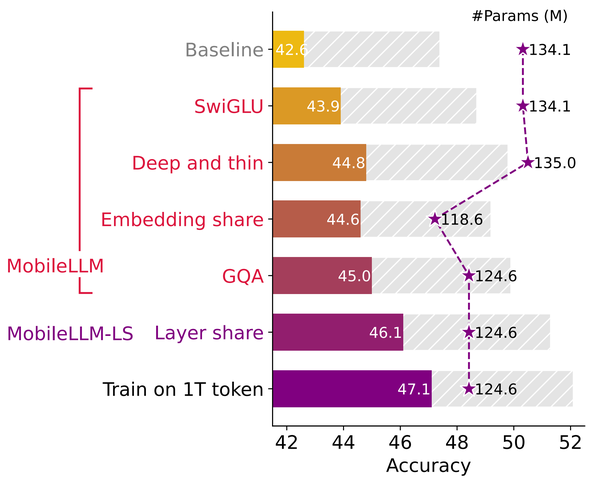
MobileLLM:
Optimizing Sub-billion Parameter Language Models for On-Device Use Cases

Zechun Liu, Changsheng Zhao, Forrest Iandola, Chen Lai, Yuandong Tian, Igor Fedorov, Yunyang Xiong,
Ernie Chang, Yangyang Shi, Raghuraman Krishnamoorthi, Liangzhen Lai, Vikas Chandra

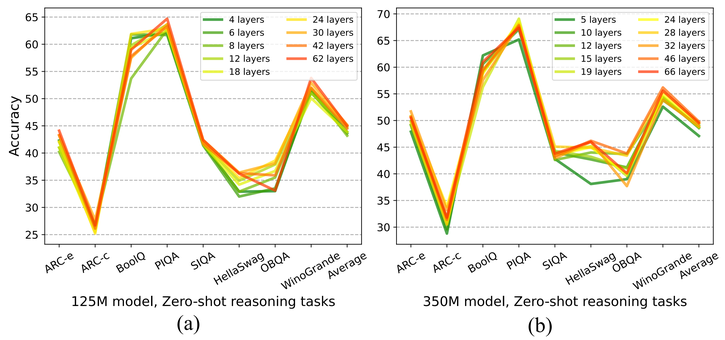
SOTA COMPARISON



DESIGN CHOICES



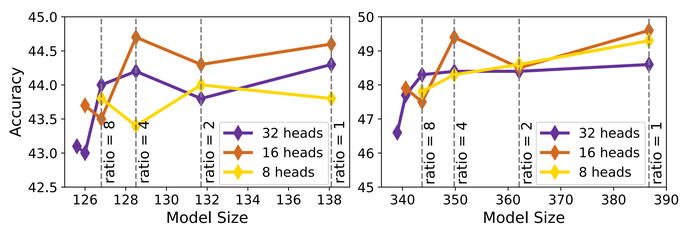
DEPTH VS WIDTH



125M model, Zero-shot reasoning tasks (a)

350M model, Zero-shot reasoning tasks (b)

Model	# Params	ARC-e	ARC-c	BoolQ	PIQA	SIQA	HS	OBQA	WinoGrande	Avg.
Without emb-share	135M	43.6	26.1	58.0	62.5	42.6	36.5	37.5	51.5	44.8
+ emb-share	119M	44.4	26.0	56.2	62.8	43.1	35.9	36.0	52.6	44.6
+ emb-share, ↑ depth	125M	43.3	26.4	54.4	64.7	43.5	36.9	38.5	52.6	45.0



(a) 125M (b) 350M

Sharing method	ARC-e	ARC-c	BoolQ	PIQA	SIQA	HellaSwag	OBQA	WinoGrande	Avg.
baseline	41.6	25.7	61.1	62.4	43.1	34.4	36.9	51.6	44.6
Immediate block-wise share	43.9	27.9	61.5	64.3	41.5	35.5	35.1	50.2	45.0

Table 7: Latency analysis of MobileLLM-125M (30 layers), MobileLLM-LS-125M (2×30 layers, adjacent blocks sharing weights), and a 60-layer non-shared weight model, with consistent configurations in all other aspects.

	Load	Init	Execute
MobileLLM	39.2 ms	1361.7 ms	15.6 ms
MobileLLM-LS	43.6 ms	1388.2 ms	16.0 ms
60-layer non-shared	68.6 ms	3347.7 ms	29.0 ms

Table 3: Zero-shot performance on Common Sense Reasoning tasks. MobileLLM denotes the proposed baseline model and MobileLLM-LS is integrated with layer sharing with the #layer counting layers with distinct weights.

Model	#Layers	#Params	ARC-e	ARC-c	BoolQ	PIQA	SIQA	HellaSwag	OBQA	WinoGrande	Avg.
Cerebras-GPT-111M	10	111M	35.8	20.2	**62.0**	58.0	39.8	26.7	29.0	48.8	40.0
LaMini-GPT-124M	12	124M	43.6	26.0	51.8	62.7	42.1	30.2	29.6	49.2	41.9
Galactica-125M	12	125M	**44.0**	26.2	54.9	55.4	38.9	29.6	28.2	49.6	40.9
OPT-125M	12	125M	41.3	25.2	57.5	62.0	41.9	31.1	31.2	50.8	42.6
GPT-neo-125M	12	125M	40.7	24.8	**61.3**	62.5	41.9	29.7	31.6	50.7	42.9
Pythia-160M	12	162M	40.0	25.3	59.5	62.0	41.5	29.9	31.2	50.9	42.5
RWKV-169M	12	169M	42.5	25.3	59.1	63.9	40.7	31.9	33.8	51.5	43.6
MobileLLM-125M	30	125M	43.9	**27.1**	60.2	**65.3**	**42.4**	**38.9**	**39.5**	**53.1**	**46.3**
MobileLLM-LS-125M	30	125M	**45.8**	**28.7**	60.4	**65.7**	**42.9**	**39.5**	**41.1**	**52.1**	**47.0**
Cerebras-GPT-256M	14	256M	37.9	23.2	60.3	61.4	40.6	28.3	31.8	50.5	41.8
OPT-350M	24	331M	41.9	25.7	54.0	64.8	42.6	36.2	33.3	52.4	43.9
Pythia-410M	24	405M	47.1	30.3	55.3	67.2	43.1	40.1	36.2	53.4	46.6
RWKV-430M	24	430M	48.9	32.0	53.4	68.1	43.6	40.6	37.8	51.6	47.0
BLOOM-560M	24	559M	43.7	27.5	53.7	65.1	42.5	36.5	32.6	52.2	44.2
Cerebras-GPT-590M	18	590M	42.6	24.9	57.7	62.8	40.9	32.0	33.2	49.7	43.0
MobileLLM-350M	32	345M	**53.8**	**33.5**	**62.4**	**68.6**	**44.7**	**49.6**	**40.0**	**57.6**	**51.3**
MobileLLM-LS-350M	32	345M	**54.4**	**32.5**	**62.8**	**69.8**	**44.1**	**50.6**	**45.8**	**57.2**	**52.1**

Table 8: Zero-shot performance on Common Sense Reasoning tasks for MobileLLM-600M, 1B and 1.5B. The highest and second-highest average scores within each model-size category are highlighted.

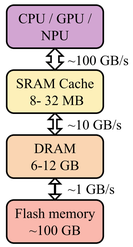
Model	ARC-e	ARC-c	BoolQ	PIQA	SIQA	HellaSwag	OBQA	WinoGrande	Avg.
Qwen1.5-500M	54.7	32.1	46.9	68.9	46.0	48.8	37.7	55.0	48.8
BLOOM-560M	43.7	27.5	53.7	65.1	42.5	36.5	32.6	52.2	44.2
Cerebras-GPT-590M	42.6	24.9	57.7	62.8	40.9	32.0	33.2	49.7	43.0
MobiLlama-800M	52.0	31.7	54.6	73.0	43.3	52.3	42.5	56.3	**50.7**
MobileLLM-600M	58.1	35.8	61.0	72.3	44.9	55.9	47.9	58.6	**54.3**
Pythia-1B	49.9	30.4	58.7	69.2	43.3	47.4	38.6	52.2	48.7
MobiLlama-1B	59.7	38.4	59.2	74.5	44.9	62.0	43.7	59.0	55.2
Falcon-1B	59.5	38.4	63.9	74.6	44.6	62.9	45.6	60.9	**56.3**
BLOOM-1.1B	47.6	27.3	58.6	67.0	42.4	42.2	36.6	53.8	46.9
TinyLlama-1.1B	59.2	37.1	58.1	72.9	43.9	59.1	44.7	58.8	54.2
MobileLLM-1B	63.0	39.0	66.7	74.4	45.0	61.4	46.8	62.3	**57.3**
Cerebras-GPT-1.3B	47.4	28.3	57.3	66.9	43.1	38.2	38.4	52.1	46.5
Galactica-1.3B	59.8	34.3	61.4	63.9	42.0	40.9	33.8	54.9	48.9
GPT-neo-1.3B	51.3	33.0	61.8	70.9	43.7	48.6	41.2	54.5	50.6
OPT-1.3B	54.4	31.7	58.4	71.5	44.7	53.7	44.6	59.1	52.3
LaMini-GPT-1.5B	59.9	39.1	77.0	71.9	45.9	50.9	44.4	57.5	55.8
RWKV-1.5B	56.2	33.8	61.8	72.3	44.7	52.8	41.8	54.7	52.3
BLOOM-1.7B	50.9	31.2	61.7	70.0	43.2	47.2	36.2	56.1	49.6
Qwen1.5-1.8B	61.1	36.5	68.3	74.1	47.2	60.4	42.9	61.2	**56.5**
Cerebras-GPT-2.7B	53.8	32.3	55.0	71.0	43.3	48.9	40.6	55.7	50.1
GPT-neo-2.7B	55.8	34.3	62.4	72.9	43.6	55.6	40.0	57.9	52.8
OPT-2.7B	56.6	34.6	61.8	74.5	45.6	60.2	48.2	59.6	55.1
Pythia-2.8B	59.4	38.9	66.1	73.8	44.5	59.6	45.0	59.4	55.8
BLOOM-3B	55.1	33.6	62.1	70.5	43.2	53.9	41.6	58.2	52.3
RWKV-3B	60.1	39.1	58.6	74.5	45.1	59.8	44.6	59.1	55.1
MobileLLM-1.5B	67.5	40.9	65.7	74.8	46.4	64.5	50.5	64.7	**59.4**

MobileLLM:
Optimizing Sub-billion Parameter Language Models for On-Device Use Cases

Zechun Liu, Changsheng Zhao, Forrest Iandola, Chen Lai, Yuandong Tian, Igor Fedorov, Yunyang Xiong,
Ernie Chang, Yangyang Shi, Raghuraman Krishnamoorthi, Liangzhen Lai, Vikas Chandra

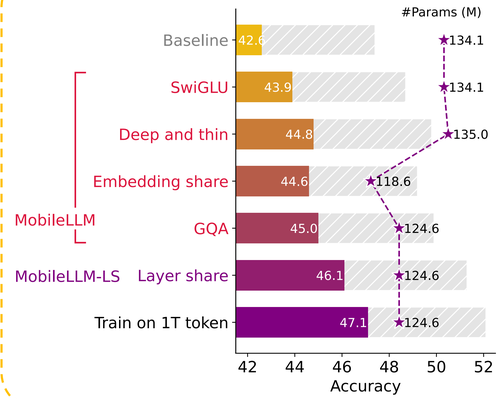
Code available:

Motivation

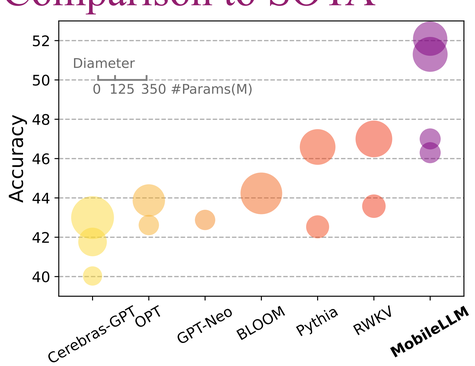


Hardware	Device	SoC last level memory size	DRAM size
Apple A16	iPhone 15	24 MB	6 GB
Apple A15	iPhone 14	32 MB	6 GB
Google	Pixel 8	8 MB	8 GB / 12 GB (pro)
QCOM	Snapdragon 8	10 MB	8-12 GB

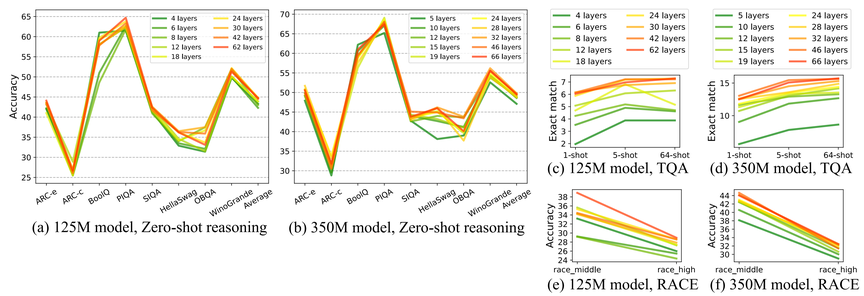
Design Choices



Comparison to SOTA



1. Depth vs Width



(a) 125M model, Zero-shot reasoning (b) 350M model, Zero-shot reasoning

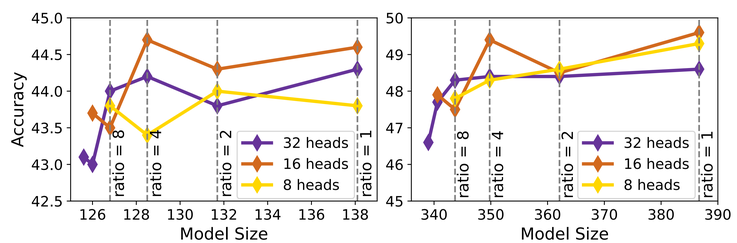
(c) 125M model, TQA (d) 350M model, TQA

(e) 125M model, RACE (f) 350M model, RACE

2. Embedding sharing

Model	# Params	ARC-e	ARC-c	BoolQ	PIQA	SIQA	HS	OBQA	WinoGrande	Avg.
Without emb-share	135M	43.6	26.1	58.0	62.5	42.6	36.5	37.5	51.5	44.8
+ emb-share	119M	44.4	26.0	56.2	62.8	43.1	35.9	36.0	52.6	44.6
+ emb-share, ↑ depth	125M	43.3	26.4	54.4	64.7	43.5	36.9	38.5	52.6	45.0

3. Group query attention



(a) 125M (b) 350M

4. Layer sharing

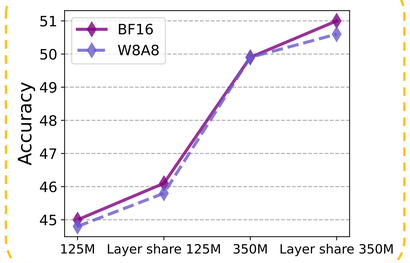
Sharing method	ARC-e	ARC-c	BoolQ	PIQA	SIQA	HellaSwag	OBQA	WinoGrande	Avg.
baseline	41.6	25.7	61.1	62.4	43.1	34.4	36.9	51.6	44.6
Immediate block-wise share	43.9	27.9	61.5	64.3	41.5	35.5	35.1	50.2	45.0
Repeat-all-over share	43.6	27.1	60.7	63.4	42.6	35.5	36.9	51.7	**45.2**
Reverse share	43.8	26.0	58.9	62.9	42.2	35.2	36.8	52.2	44.8
baseline	50.8	30.6	62.3	68.6	43.5	45.1	43.8	52.4	49.6
Immediate block-wise share	51.5	30.8	59.6	68.2	43.9	47.7	44.7	55.0	50.2
Repeat-all-over share	53.5	33.0	61.2	69.4	43.2	48.3	42.2	54.6	**50.7**
Reverse share	50.7	32.2	61.0	68.8	43.8	47.4	43.1	53.8	50.1

Latency

Table 7: Latency analysis of MobileLLM-125M (30 layers), MobileLLM-LS-125M (2×30 layers, adjacent blocks sharing weights), and a 60-layer non-shared weight model, with consistent configurations in all other aspects.

	Load	Init	Execute
MobileLLM	39.2 ms	1361.7 ms	15.6 ms
MobileLLM-LS	43.6 ms	1388.2 ms	16.0 ms
60-layer non-shared	68.6 ms	3347.7 ms	29.0 ms

Quantization



Final results

Model	#Layers	#Params	ARC-e	ARC-c	BoolQ	PIQA	SIQA	HellaSwag	OBQA	WinoGrande	Avg.
Cerebras-GPT-111M	10	111M	35.8	20.2	**62.0**	58.0	39.8	26.7	29.0	48.8	40.0
LaMini-GPT-124M	12	124M	43.6	26.0	51.8	62.7	42.1	30.2	29.6	49.2	41.9
Galactica-125M	12	125M	**44.0**	26.2	54.9	55.4	38.9	29.6	28.2	49.6	40.9
OPT-125M	12	125M	41.3	25.2	57.5	62.0	41.9	31.1	31.2	50.8	42.6
GPT-neo-125M	12	125M	40.7	24.8	**61.3**	62.5	41.9	29.7	31.6	50.7	42.9
Pythia-160M	12	162M	40.0	25.3	59.5	62.0	41.5	29.9	31.2	50.9	42.5
RWKV-169M	12	169M	42.5	25.3	59.1	63.9	40.7	31.9	33.8	51.5	43.6
MobileLLM-125M	30	125M	43.9	**27.1**	60.2	**65.3**	**42.4**	**38.9**	**39.5**	**53.1**	**46.3**
MobileLLM-LS-125M	30	125M	**45.8**	**28.7**	60.4	**65.7**	**42.9**	**39.5**	**41.1**	**52.1**	**47.0**
Cerebras-GPT-256M	14	256M	37.9	23.2	60.3	61.4	40.6	28.3	31.8	50.5	41.8
OPT-350M	24	331M	41.9	25.7	54.0	64.8	42.6	36.2	33.3	52.4	43.9
Pythia-410M	24	405M	47.1	30.3	55.3	67.2	43.1	40.1	36.2	53.4	46.6
RWKV-430M	24	430M	48.9	32.0	53.4	68.1	43.6	40.6	37.8	51.6	47.0
BLOOM-560M	24	559M	43.7	27.5	53.7	65.1	42.5	36.5	32.6	52.2	44.2
Cerebras-GPT-590M	18	590M	42.6	24.9	57.7	62.8	40.9	32.0	33.2	49.7	43.0
MobileLLM-350M	32	345M	**53.8**	**33.5**	**62.4**	**68.6**	**44.7**	**49.6**	**40.0**	**57.6**	**51.3**
MobileLLM-LS-350M	32	345M	**54.4**	**32.5**	**62.8**	**69.8**	**44.1**	**50.6**	**45.8**	**57.2**	**52.1**

Model	ARC-e	ARC-c	BoolQ	PIQA	SIQA	HellaSwag	OBQA	WinoGrande	Avg.
Qwen1.5-500M	54.7	32.1	46.9	68.9	46.0	48.8	37.7	55.0	48.8
BLOOM-560M	43.7	27.5	53.7	65.1	42.5	36.5	32.6	52.2	44.2
Cerebras-GPT-590M	42.6	24.9	57.7	62.8	40.9	32.0	33.2	49.7	43.0
MobiLlama-800M	52.0	31.7	54.6	73.0	43.3	52.3	42.5	56.3	**50.7**
MobileLLM-600M	58.1	35.8	61.0	72.3	44.9	55.9	47.9	58.6	**54.3**
Pythia-1B	49.9	30.4	58.7	69.2	43.3	47.4	38.6	52.2	48.7
MobiLlama-1B	59.7	38.4	59.2	74.5	44.9	62.0	43.7	59.0	55.2
Falcon-1B	59.5	38.4	63.9	74.6	44.6	62.9	45.6	60.9	**56.3**
BLOOM-1.1B	47.6	27.3	58.6	67.0	42.4	42.2	36.6	53.8	46.9
TinyLlama-1.1B	59.2	37.1	58.1	72.9	43.9	59.1	44.7	58.8	54.2
MobileLLM-1B	63.0	39.0	66.7	74.4	45.0	61.4	46.8	62.3	**57.3**
Cerebras-GPT-1.3B	47.4	28.3	57.3	66.9	43.1	38.2	38.4	52.1	46.5
Galactica-1.3B	59.8	34.3	61.4	63.9	42.0	40.9	33.8	54.9	48.9
GPT-neo-1.3B	51.3	33.0	61.8	70.9	43.7	48.6	41.2	54.5	50.6
OPT-1.3B	54.4	31.7	58.4	71.5	44.7	53.7	44.6	59.1	52.3
LaMini-GPT-1.5B	59.9	39.1	77.0	71.9	45.9	50.9	44.4	57.5	55.8
RWKV-1.5B	56.2	33.8	61.8	72.3	44.7	52.8	41.8	54.7	52.3
BLOOM-1.7B	50.9	31.2	61.7	70.0	43.2	47.2	36.2	56.1	49.6
Qwen1.5-1.8B	61.1	36.5	68.3	74.1	47.2	60.4	42.9	61.2	**56.5**
Cerebras-GPT-2.7B	53.8	32.3	55.0	71.0	43.3	48.9	40.6	55.7	50.1
GPT-neo-2.7B	55.8	34.3	62.4	72.9	43.6	55.6	40.0	57.9	52.8
OPT-2.7B	56.6	34.6	61.8	74.5	45.6	60.2	48.2	59.6	55.1
Pythia-2.8B	59.4	38.9	66.1	73.8	44.5	59.6	45.0	59.4	55.8
BLOOM-3B	55.1	33.6	62.1	70.5	43.2	53.9	41.6	58.2	52.3
RWKV-3B	60.1	39.1	58.6	74.5	45.1	59.8	44.6	59.1	55.1
MobileLLM-1.5B	67.5	40.9	65.7	74.8	46.4	64.5	50.5	64.7	**59.4**