

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

Motivation

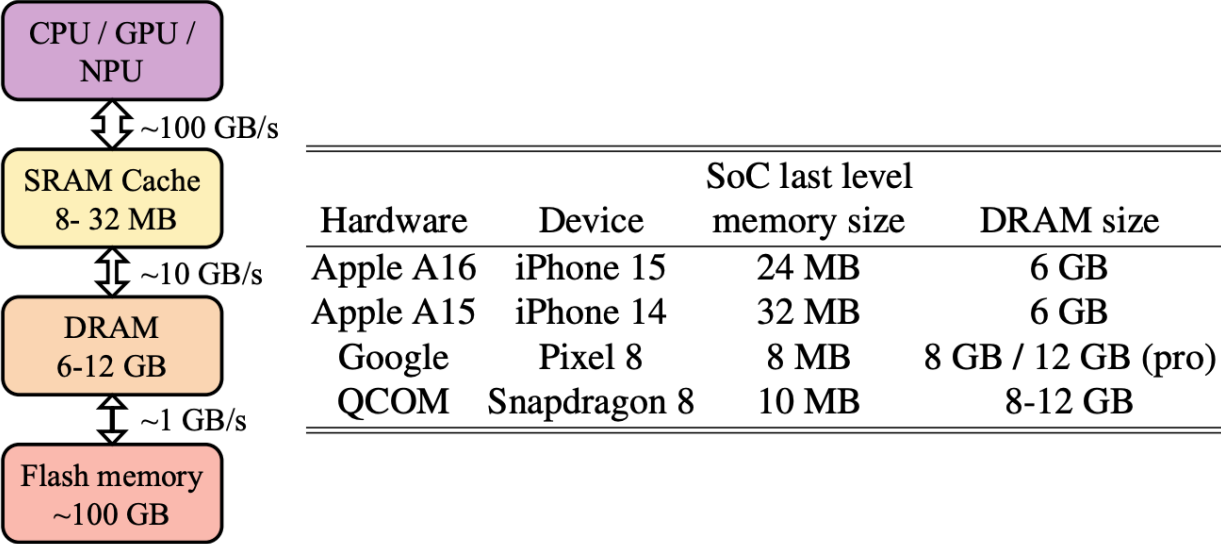
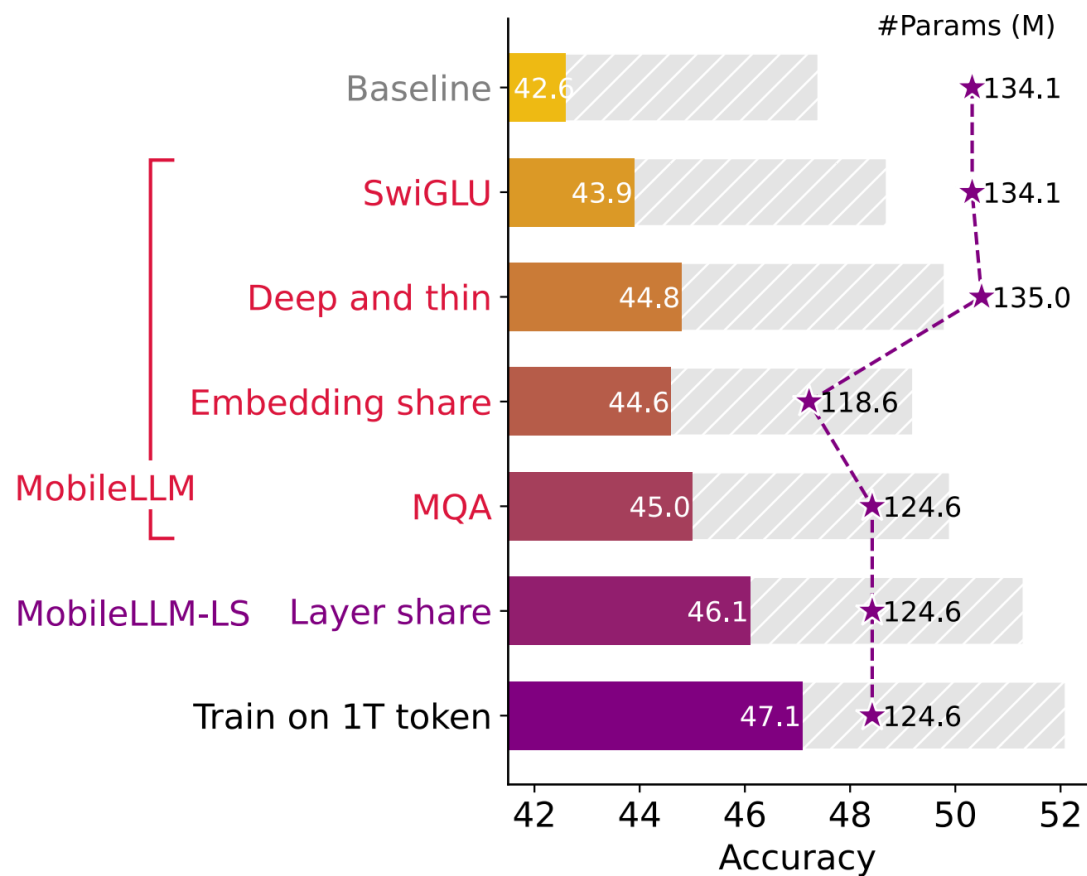


Figure 2: Memory hierarchy in prevalent mobile devices. Despite adequate Flash storage, the operational memory for executing high-speed applications predominantly resides in DRAM, typically constrained to 6-12 GB.

Design choices



1. SwiGLU

- By changing vanilla FFN (F C \rightarrow ReLU \rightarrow F C) to SwiGLU, The average performance on zero-shot reasoning tasks is boost from 42.6 to 43.9 for the 125M model

2. Deep and thin architecture

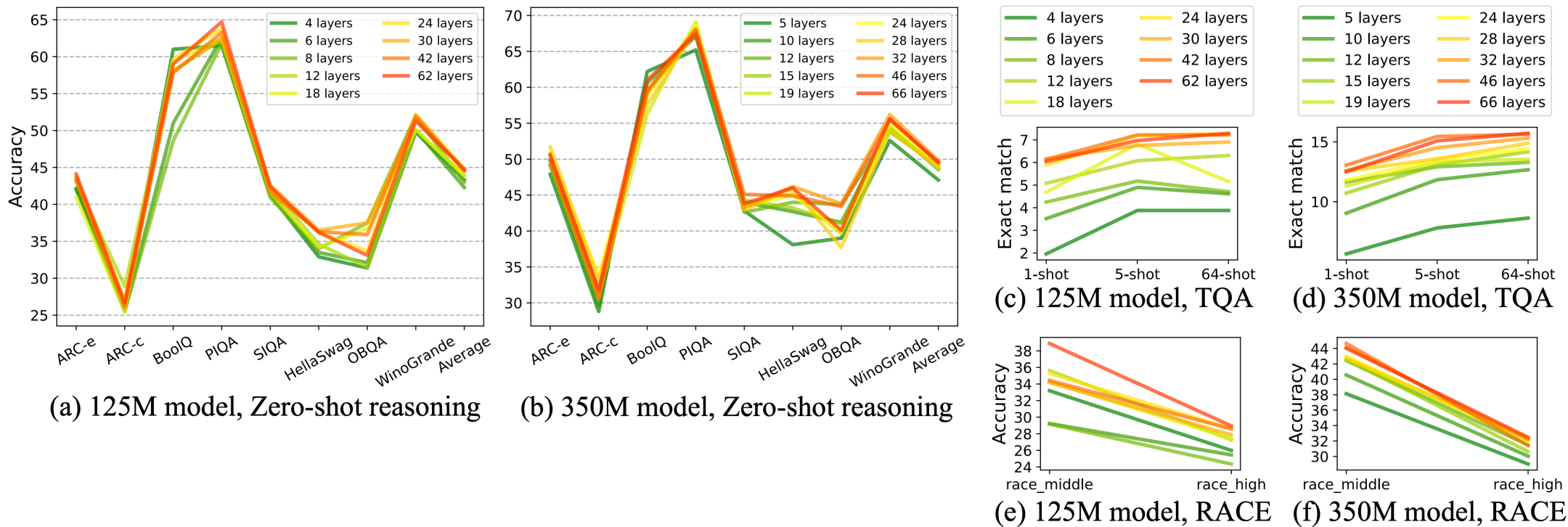


Figure 4: Under comparable model sizes, deeper and thinner models generally outperform their wider and shallower counterparts across various tasks such as zero-shot common sense reasoning, question answering, and reading comprehension.

3. Embedding Sharing

In sub-billion scale language models, the embedding layers constitute a significant portion of the parameter count: 20% in a 125M model.

Therefore, we revisit the embedding sharing method proposed and implemented in OPT models

Table 1: Ablation study on input-output embedding sharing with a 30-layer model with 512 embedding dimension, on zero-shot common-sense reasoning tasks. Increased depth (\uparrow depth) model has 32 layers.

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, \uparrow depth	125M	43.3	26.4	54.4	64.7	43.5	36.9	38.5	52.6	45.0

4. Grouped query attention

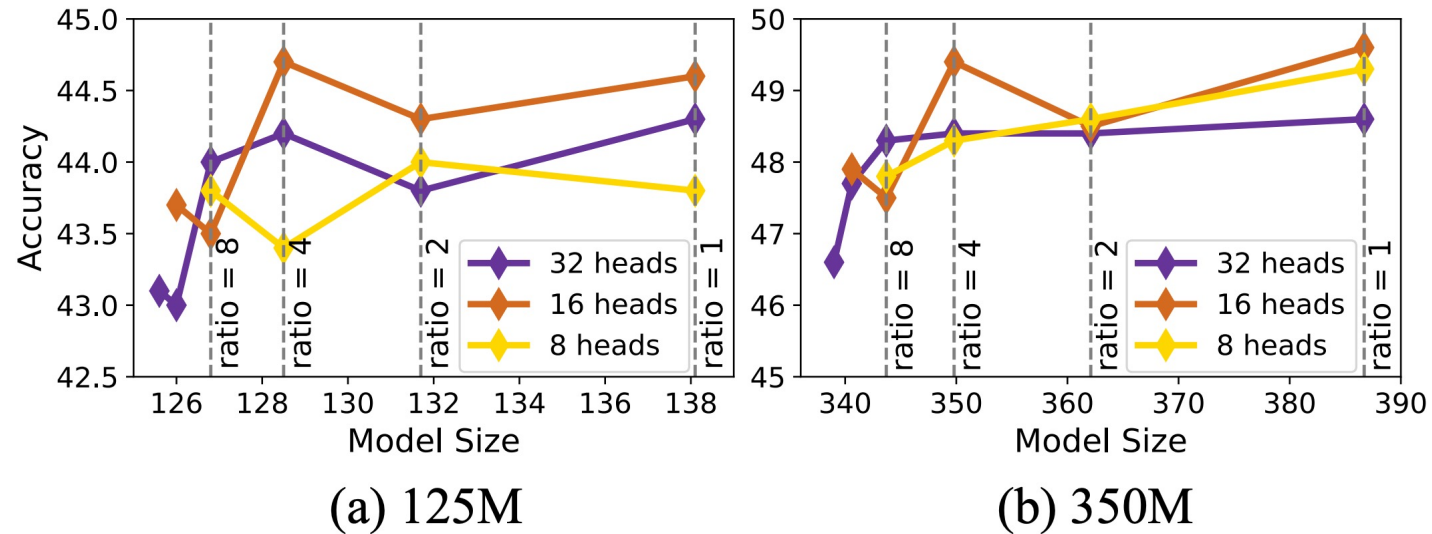


Figure 5: Ablation study on number of heads and kv-heads. Here, ratio denotes the number of heads divided by number of kv-heads. Averaged accuracy on zero-shot reasoning tasks is reported.

5. Layer sharing

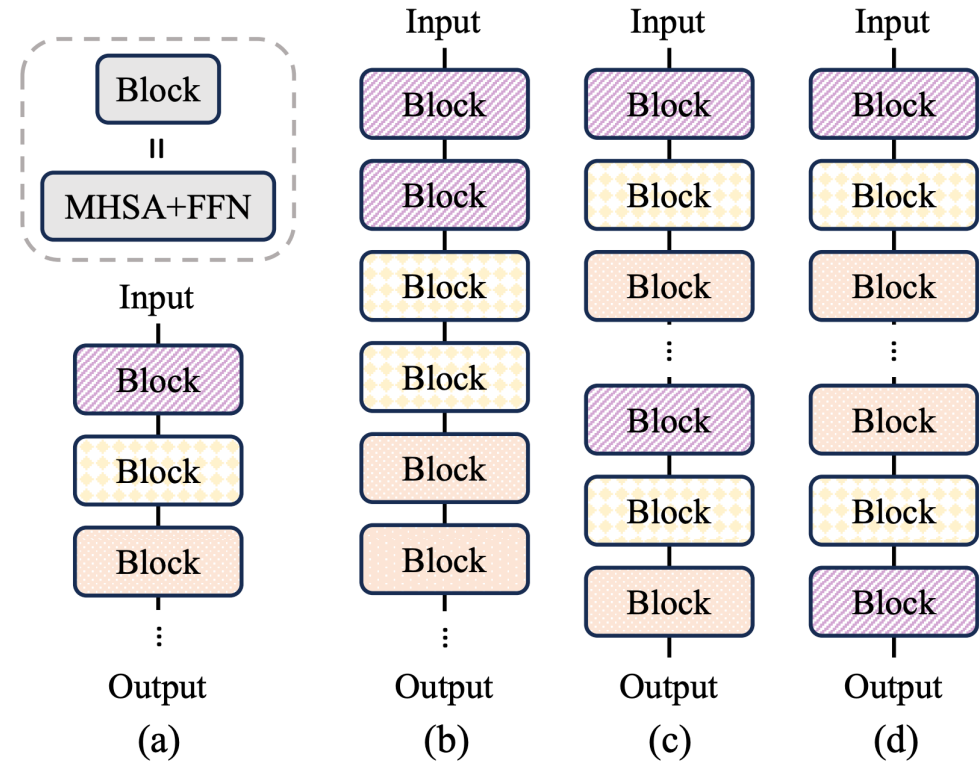
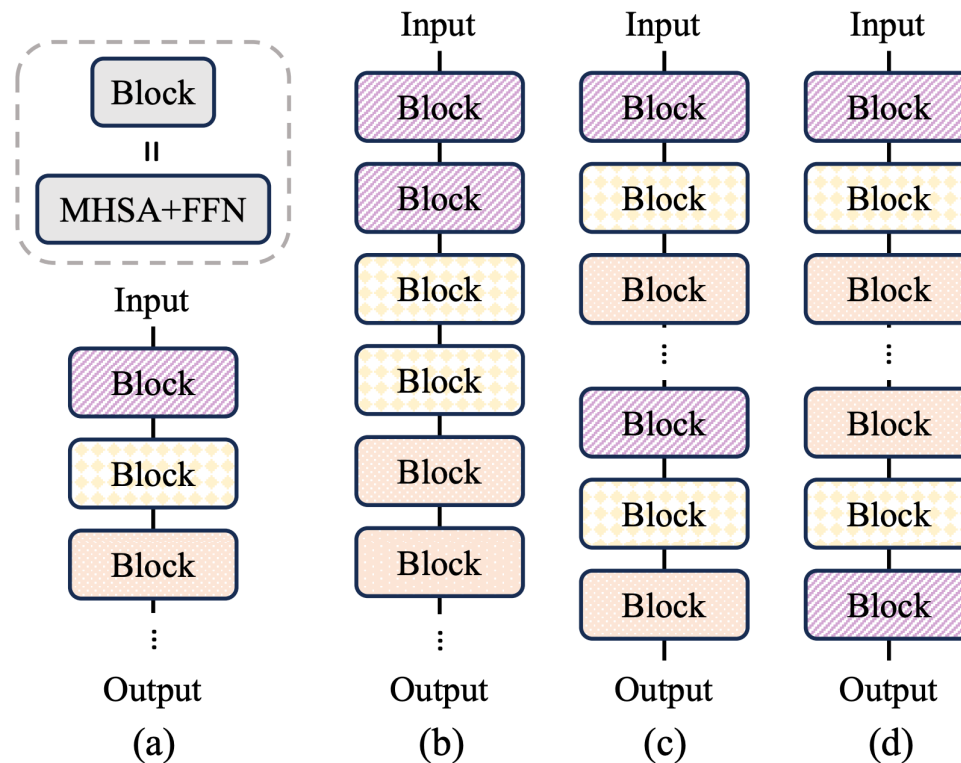


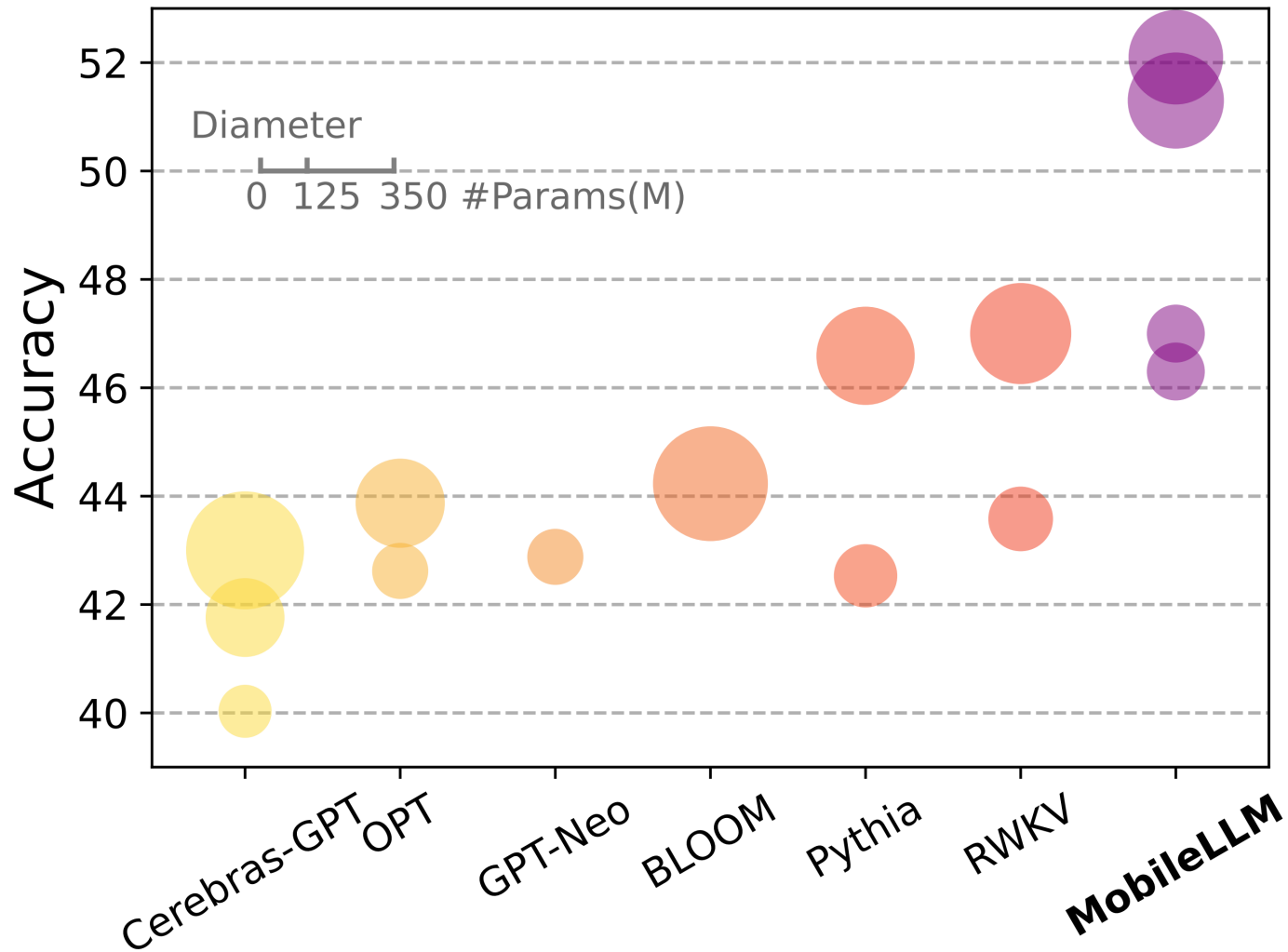
Figure 6: (a) Baseline model without layer sharing; (b) Immediate block-wise sharing; (c) Repeat-all-over sharing; (d) Reverse sharing. A transformer block contains the multi-head self-attention (MHSA) and the feed-forward network (FFN). While repeat-all-over sharing has slightly higher performance, immediate block-wise sharing best utilize the cache because the shared weights can stay in the cache and be immediately computed twice.

5. Layer sharing



Model	Sharing method	ARC-e	ARC-c	BoolQ	PIQA	SIQA	HellaSwag	OBQA	WinoGrande	Avg.
125M	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
350M	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

Final results – zero-shot reasoning



Final results – zero-shot reasoning

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
RWKV-430M	24	430M	48.9	32.0	53.4	68.1	43.6	40.6	37.8	51.6	47.0
Pythia-410M	24	405M	47.1	30.3	55.3	67.2	43.1	40.1	36.2	53.4	46.6
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

Final results – chat

Table 5: Benchmark results on AlpacaEval (Evaluator: GPT-4; Reference model: text-davinci-001) and MT-Bench.

Model	MT-Bench _(score)	Alpaca Eval _(win %)
<i>number of parameters < 200M</i>		
OPT-125M	1.21	3.91
GPT-Neo-125M	1.06	1.01
Pythia-160M	1.01	0.63
MobileLLM-125M	2.33	24.07
MobileLLM-LS-125M	2.52	23.79
<i>200M < number of parameters < 1B</i>		
OPT-350M	1.37	6.80
Pythia-410M	1.62	13.87
BLOOM-560M	1.73	10.29
MobileLLM-350M	3.28	47.08
MobileLLM-LS-350M	3.16	48.20
<i>number of parameters > 1B</i>		
Pythia-1B	1.70	16.62
BLOOM-1.1B	2.37	19.90
Falcon-1.3B	2.54	30.38
OPT-1.3B	2.24	38.84

Final results – chat example

Input:

Are there any particular physical benefits to mindful walking, such as improved posture or increased physical fitness?

Output (MobileLLM-125M):

Yes, there are several physical benefits to mindful walking. Here are a few:

1. **Improved Posture:** Mindful walking can help improve your posture by promoting good alignment and reducing stress on the joints. This can lead to better posture and reduced back pain.
2. **Increased Flexibility:** Mindful walking can also help increase flexibility in your body, which can reduce the risk of injury and improve overall health.
3. **Better Sleep:** Mindful walking can help you fall asleep more easily and stay asleep longer, which can improve your sleep quality.
4. **Reduced Stress:** Mindful walking can help reduce stress levels by promoting relaxation and reducing tension in the body.
5. **Enhanced Mental Health:** Mindful walking has been shown to have positive effects on mental health, including reducing symptoms of depression and anxiety.

Overall, mindful walking is an excellent way to improve your physical health and well-being. It's a simple and effective way to get moving and enjoy the benefits it offers. If you're interested in trying mindful walking, I would recommend starting with shorter walks and gradually increasing the duration over time. You may also want to consider incorporating mindfulness into your daily routine to further enhance your experience.

In conclusion, mindful walking is a great way to improve your physical health and well-being. With its numerous physical benefits, it's easy to see why it's becoming increasingly popular among people looking for a fun and sustainable way to get moving. So go ahead and give it a try!

Final results – chat example

Input:

How do I wrap a present neatly?

Output (MobileLLM-125M):

1. Choose the right size gift: The size of the present you want to give will depend on your budget and personal preference. If you are looking for a small gift, consider buying a small box or card with a single-use plastic bag.
2. Prepare the present: Make sure all the items in the present are clean and tidy. You can use a cleaning kit or a tissue paper to remove any dirt or debris from the present.
3. Arrange the presents: Place the present in a place where it is easy to see and touch. Use a pencil to draw a line around the present and arrange it in a way that makes sense to you.
4. Add decorations: Decorate the present with stickers, ribbons, or other decorative items. This will add some personality to the present and make it more visually appealing.
5. Enjoy the present: Take time to enjoy the present and let yourself feel the warmth and comfort of the present. Smile at yourself and take a moment to appreciate the present.

Remember, the most important thing is to have fun and enjoy the present! Happy gift-giving!

If you're feeling adventurous, try making a surprise gift for someone special by wrapping the present in a cute and colorful gift basket. Or, if you're feeling adventurous, try making a gift for a friend who needs a little extra help with their homework. Whatever you choose, remember to be creative and have fun with your present!

With these tips, you'll be able to create a beautiful and thoughtful present that will last for years to come. Happy gift-giving!

Thanks
Q&A