PennState

Small Language Models in the Era of Large Language Models

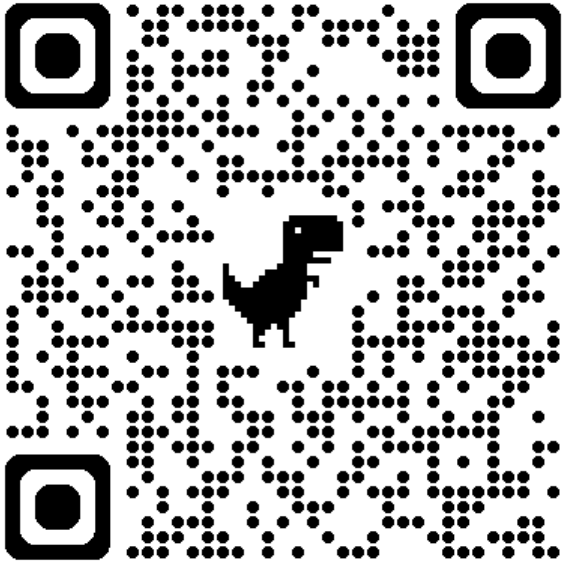
Techniques, Applications, and SLMs for LLMs

Fali Wang (fairyfali.github.io)

April 27, 2025 *(Work in submission)*

Related Materials

- Paper: arXiv
- Github: Github
- English Blog: in Linkin
- Chinese Blog: in Wechat
- Slides are in my personal page (fairyfali.github.io).



GitHub Repo

Outline[1]

- Introduction
- *Enhancement Strategy for SLMs
- *On-device SLMs and Applications
- *SLMs for LLMs
- Future Directions

[1]Here * means the section is key

Why SLMs?

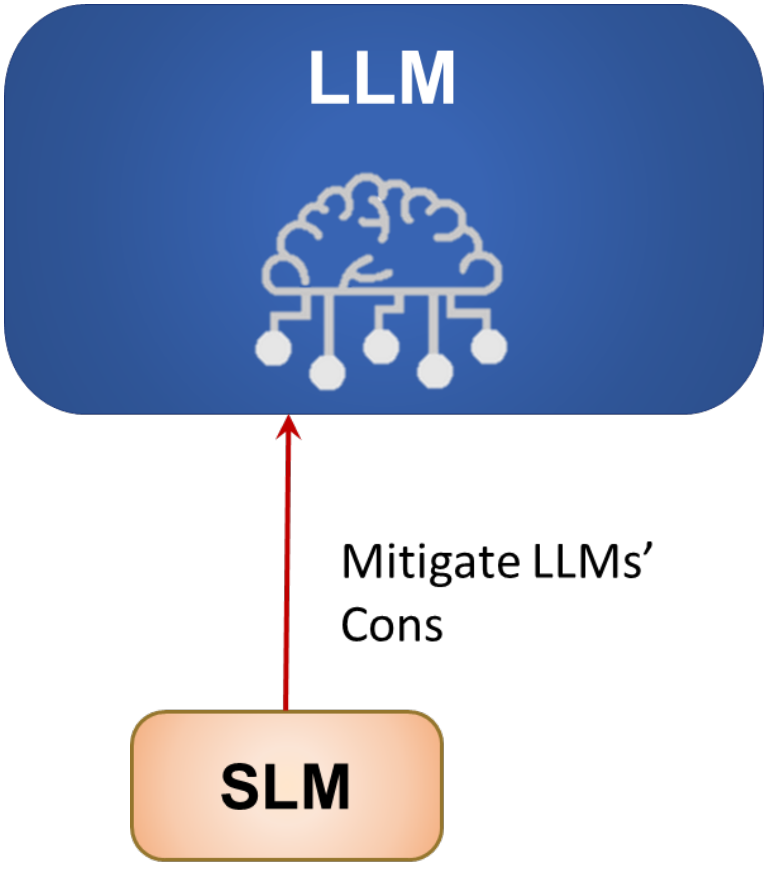
Pros:

- Emergent ability
- Generalizability

Cons:

- On-device deployment
- Privacy leakage
- Inference latency
- Expensive fine-tuning
- Inferior to specialized models

Why SLMs?



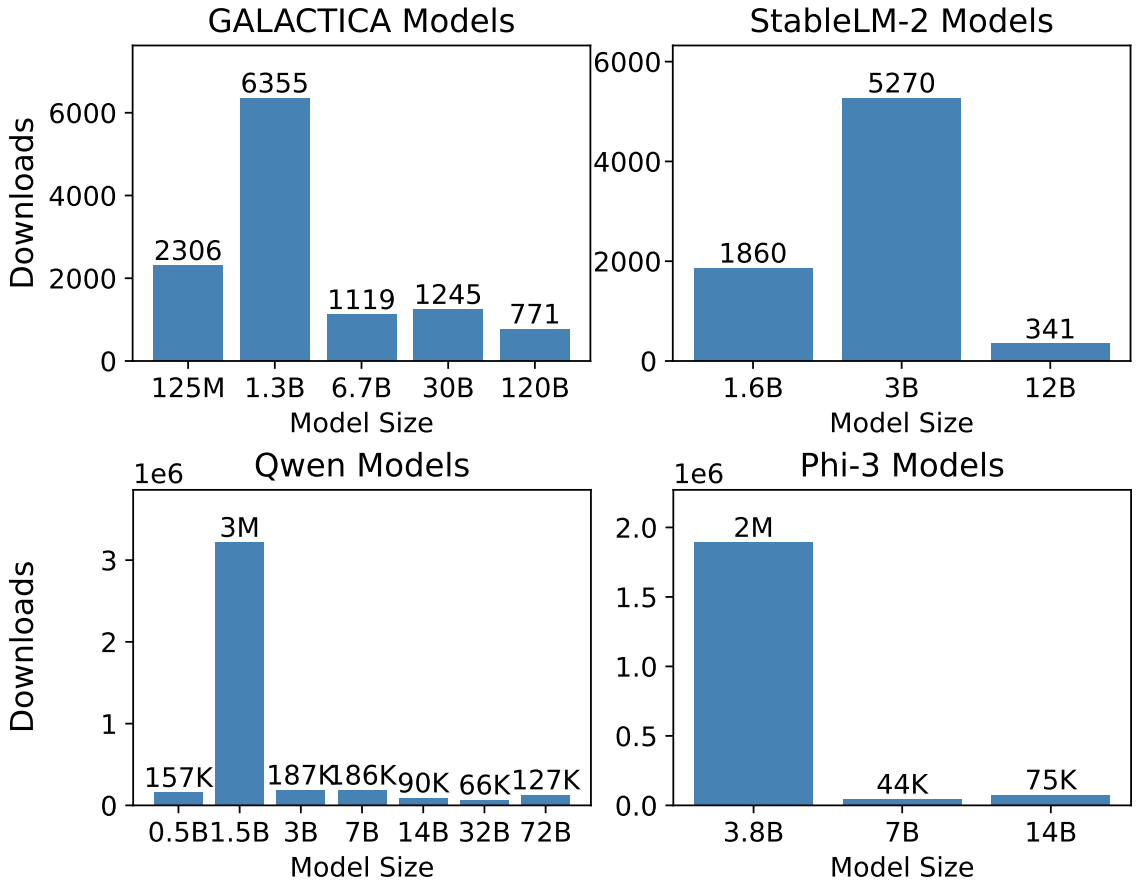
Pros:

- Emergent ability
- Generalizability

Cons:

- On-device deployment
- Privacy leakage
- Inference latency
- Expensive fine-tuning
- Inferior to specialized models

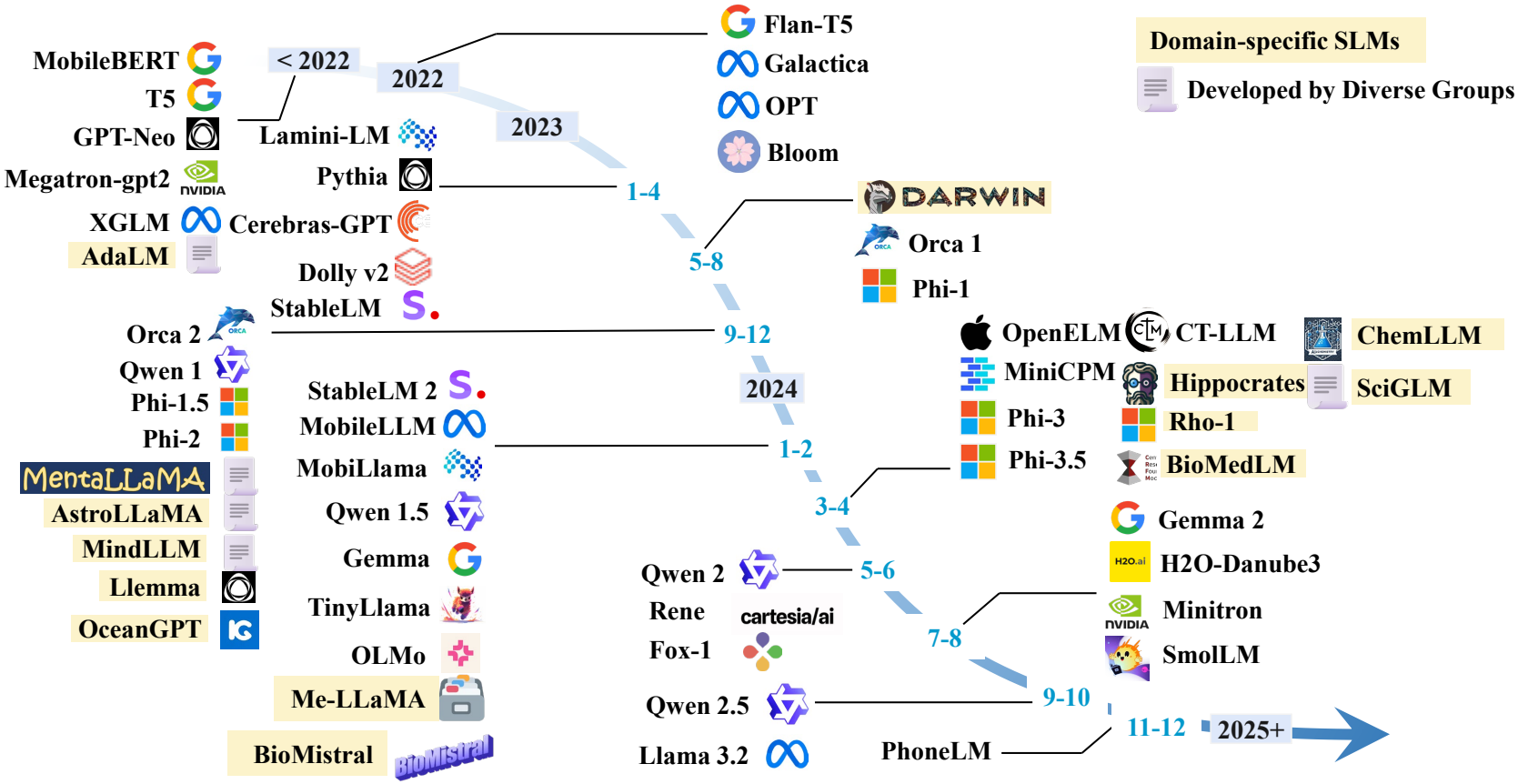
Smaller Language Models are Popular



Download Statistics obtained on October 7, 2024.

Smaller language models are downloaded more frequently than larger models in the Hugging Face community.

Timeline of Existing SLMs



Issues of Existing SLM Definition

- **Relative Definition**: Lu et al. [2024], Van Nguyen et al. [2024], Chen and Varoquaux [2024] view "small" as relative to "large."
- **Perspective of Mobile Devices**: MobileLLM Liu et al. [2024] categorizes SLMs as models with fewer than one billion parameters, suitable for mobile devices with up to 6GB memory.
- **Perspective of Emergent Ability**: SLMs typically range from a few million to a few billion (under 7B or 10B) [2], often lacking emergent abilities Fu et al. [2023].
- However, they lack consensus and no clear boundaries between SLMs and LLMs. 7B LMs belong to an LLM or SLM?

[2] The Rise of Small Language Models: Efficiency and Customization for AI

Our SLM Definition

- Considering both capability and resource constraints, our definition is:

Def 1: Our SLM Definition

Given specific tasks and resource constraints, we define SLMs as falling within a range where the lower bound is the minimum size at which the model exhibits emergent abilities for a specialized task, and the upper bound is the largest size manageable within limited resource conditions.

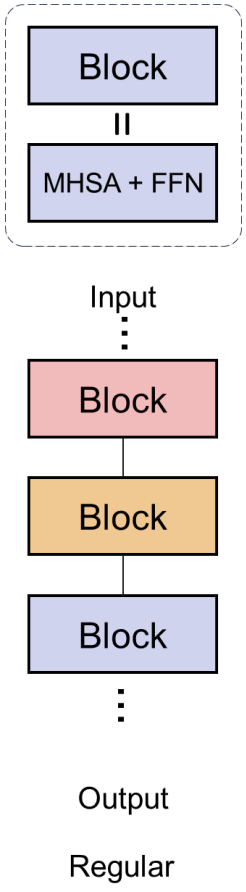
Outline

- Introduction
- *Enhancement Strategy for SLMs
- *On-device SLMs and Applications
- *SLMs for LLMs
- Future Directions

Enhancement Strategy

- **Pre-training SLMs from scratch**: Architecture choice; Parameter Sharing; Data quality.
- **Supervised Fine-tuning**: Pretrain-then-finetune; Instruction tuning; Preference optimization.

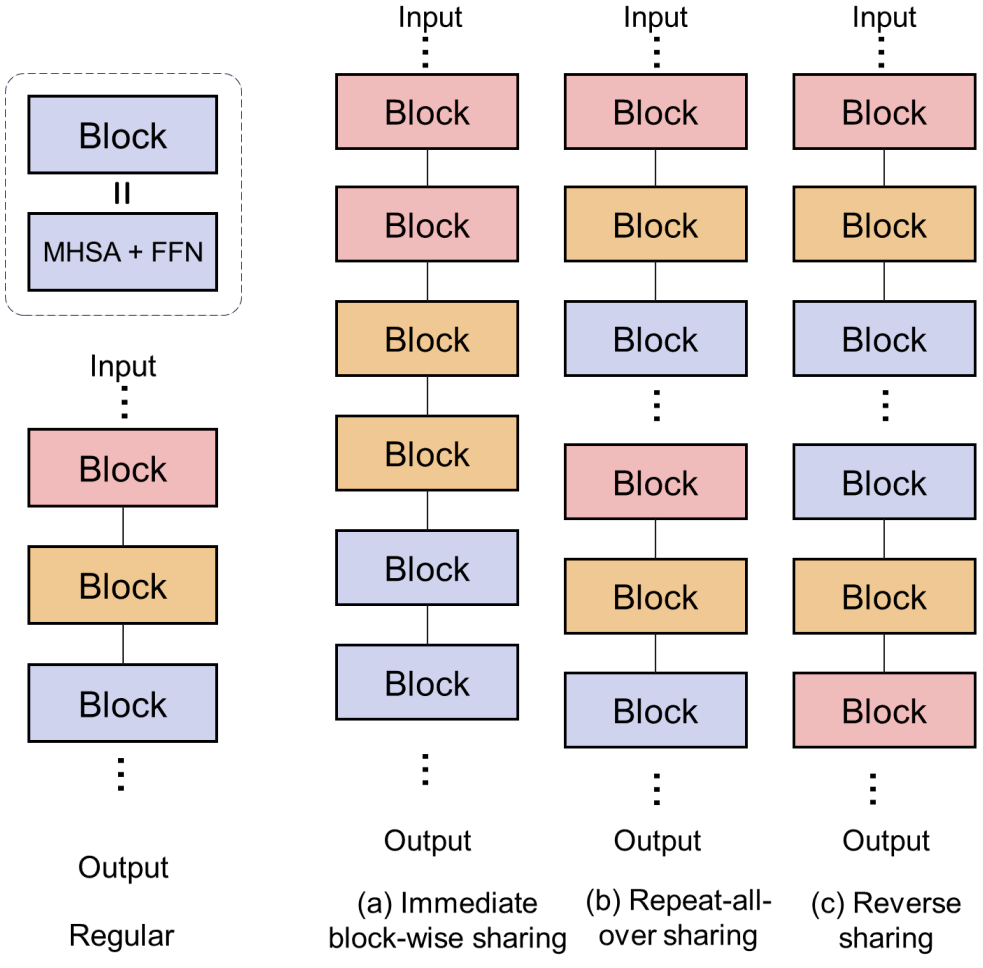
Pre-training from scratch-Parameter Sharing [3] [4]



[3] Omkar Thawakar, et al. Mobillama: Towards accurate and lightweight fully transparent GPT

[4] Zechun Liu et al., MobileLLM: Optimizing Sub-billion Parameter Language Models for On-Device Use Cases

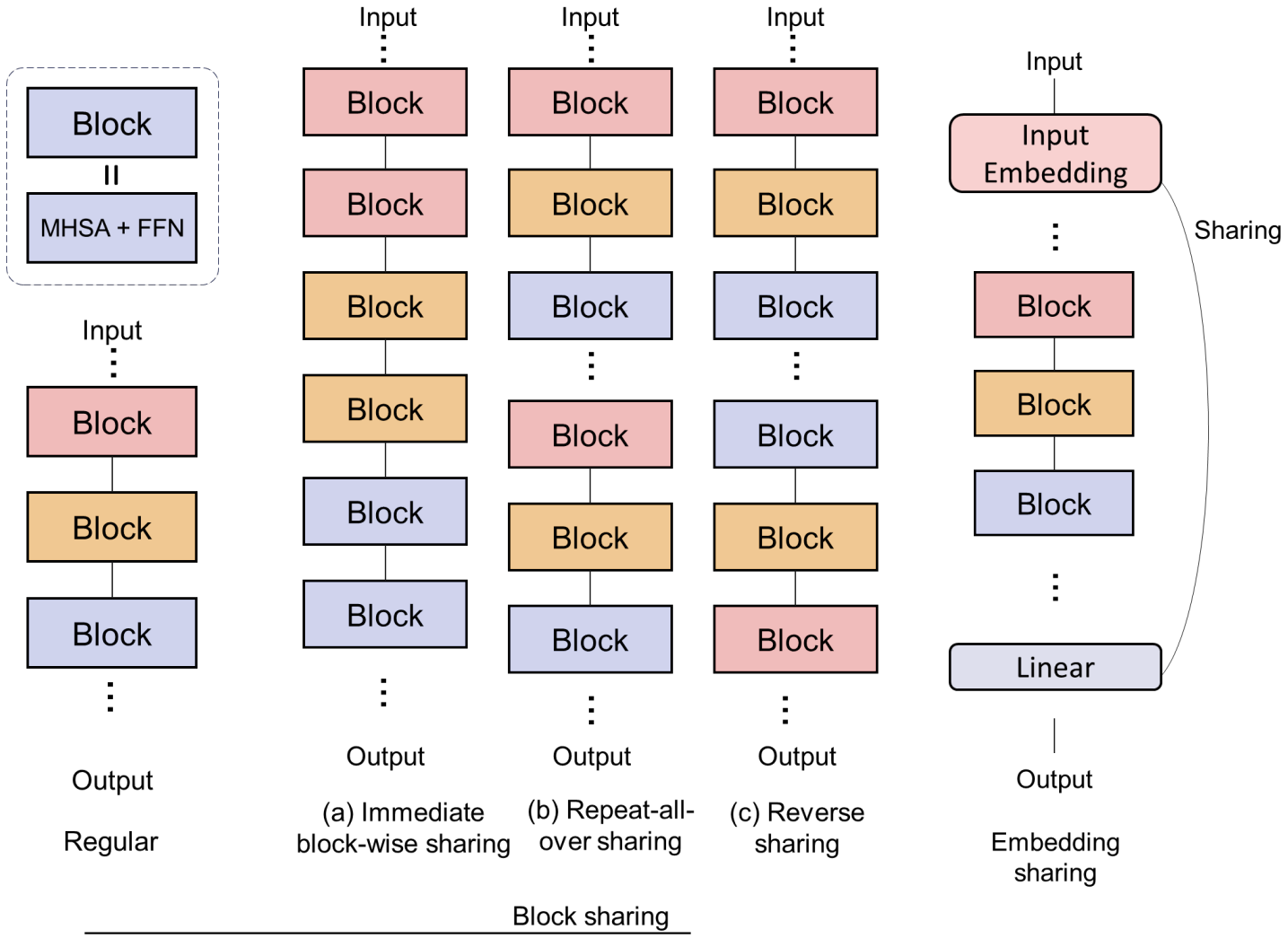
Pre-training from scratch-Parameter Sharing [3] [4]



[3] Omkar Thawakar, et al. Mobillama: Towards accurate and lightweight fully transparent GPT

[4] Zechun Liu et al., MobileLLM: Optimizing Sub-billion Parameter Language Models for On-Device Use Cases

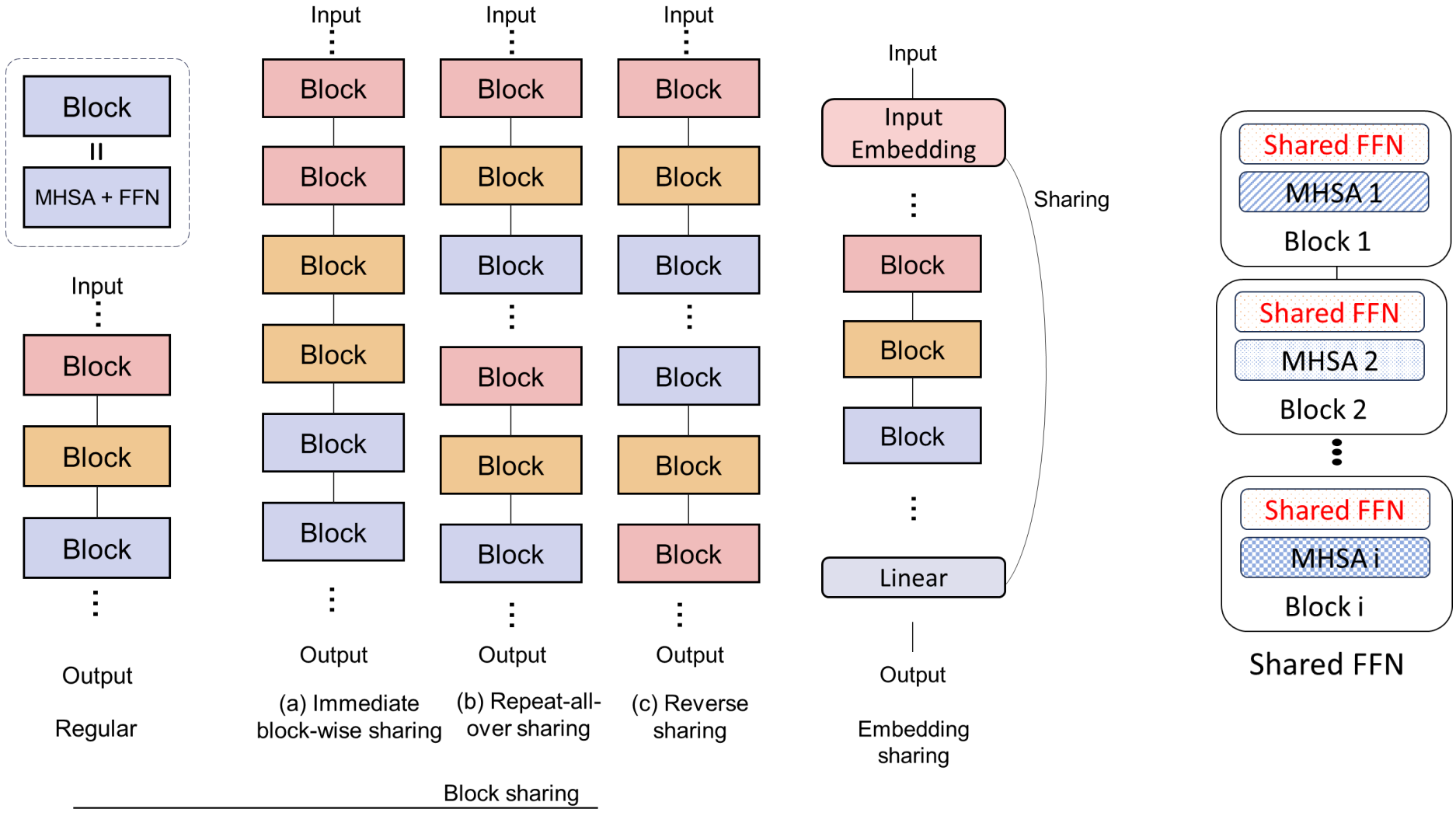
Pre-training from scratch-Parameter Sharing [3] [4]



[3] Omkar Thawakar, et al. Mobillama: Towards accurate and lightweight fully transparent GPT

[4] Zechun Liu et al., MobileLLM: Optimizing Sub-billion Parameter Language Models for On-Device Use Cases

Pre-training from scratch-Parameter Sharing [3] [4]



[3] Omkar Thawakar, et al. Mobillama: Towards accurate and lightweight fully transparent GPT

[4] Zechun Liu et al., MobileLLM: Optimizing Sub-billion Parameter Language Models for On-Device Use Cases

Data Quality in KD - TinyStories [5]

TinyStories Dataset Creation:

- Vocabulary of 1500 basic words, similar to a 3-4 year-old's vocabulary.
- Stories generated by ChatGPT/GPT-4 using three randomly selected words (a verb, a noun, and an adjective) and incorporating features like dialogue, plot twists, or morals.

TinyStories Instruction Dataset Creation:

[5] Ronen Eldan and Yuanzhi Li. TinyStories: How small can language models be and still speak coherent English?

Data Quality in KD - TinyStories [5]

Hidden size	Layer	Eval loss	Creativity	Grammar	Consistency	Instruct	Plot
64	12	2.02	4.84/0.36	6.19/0.42	4.75/0.31	4.34/0.23	4.39/0.20
64	8	2.08	4.68/0.33	6.14/0.41	4.45/0.27	4.34/0.23	4.40/0.21
64	4	2.26	3.97/0.20	5.31/0.22	3.77/0.18	3.79/0.14	3.71/0.06
64	2	2.38	2.94/0.00	4.33/0.00	2.41/0.00	2.86/0.00	3.40/0.00
128	12	1.62	6.02/0.58	7.25/0.66	7.20/0.64	6.94/0.63	6.58/0.65
128	8	1.65	5.97/0.57	7.23/0.66	7.10/0.62	6.87/0.62	6.16/0.57
128	4	1.78	5.70/0.52	6.91/0.58	6.60/0.56	6.00/0.49	5.53/0.44
128	2	1.92	4.90/0.37	6.43/0.48	4.75/0.31	5.23/0.37	4.89/0.31
256	12	1.34	6.66/0.71	7.80/0.79	8.38/0.79	7.68/0.75	7.18/0.78
256	8	1.38	6.54/0.68	7.72/0.77	8.02/0.75	7.92/0.78	7.23/0.79
256	4	1.47	6.32/0.64	7.64/0.75	7.76/0.71	8.07/0.81	7.18/0.78
256	2	1.60	6.23/0.62	7.50/0.72	7.20/0.64	7.23/0.68	6.50/0.64
512	12	1.19	6.90/0.75	8.46/0.93	9.11/0.89	8.21/0.83	7.37/0.82
512	8	1.20	6.85/0.74	8.34/0.91	8.95/0.87	8.05/0.80	7.26/0.79
512	4	1.27	6.75/0.72	8.35/0.91	8.50/0.81	8.34/0.85	7.36/0.81
512	2	1.39	6.40/0.66	7.72/0.77	7.90/0.73	7.76/0.76	7.13/0.77
768	12	1.18	7.00/0.77	8.30/0.90	9.20/0.90	8.23/0.83	7.47/0.84
768	8	1.18	7.02/0.77	8.62/0.97	9.34/0.92	8.36/0.85	7.34/0.81
768	4	1.20	6.89/0.75	8.43/0.93	9.01/0.88	8.44/0.87	7.52/0.85
768	2	1.31	6.68/0.71	8.01/0.83	8.42/0.80	7.97/0.79	7.34/0.81
768	1	1.54	6.00/0.58	7.35/0.68	7.25/0.64	5.81/0.46	6.44/0.63
1024	12	1.22	7.05/0.78	8.43/0.93	8.98/0.87	8.18/0.82	7.29/0.80
1024	8	1.20	7.13/0.80	8.25/0.89	8.92/0.87	8.47/0.87	7.47/0.84
1024	4	1.21	7.04/0.78	8.32/0.90	8.93/0.87	8.34/0.85	7.47/0.84
1024	2	1.27	6.68/0.71	8.22/0.88	8.52/0.81	8.04/0.80	7.24/0.79
1024	1	1.49	6.36/0.65	7.77/0.78	7.47/0.67	6.09/0.50	6.42/0.62
GPT-Neo (125M)	-	-	3.34/0.08	5.27/0.21	4.22/0.24	-	-
GPT-2-small (125M)	-	-	3.70/0.14	5.40/0.24	4.32/0.25	-	-
GPT-2-med (355M)	-	-	4.22/0.24	6.27/0.44	5.34/0.39	-	-
GPT-2-large (774M)	-	-	4.30/0.26	6.43/0.48	6.04/0.48	-	-
GPT-4	-	-	8.21/1.00	8.75/1.00	9.93/1.00	9.31/1.00	8.26/1.00

Data Quality in KD - TinyStories [5]

Hidden size	Layer	Eval loss	Creativity	Grammar	Consistency	Instruct	Plot
64	12	2.02	4.84/0.36	6.19/0.42	4.75/0.31	4.34/0.23	4.39/0.20
64	8	2.08	4.68/0.33	6.14/0.41	4.45/0.27	4.34/0.23	4.40/0.21
64	4	2.26	3.97/0.20	5.31/0.22	3.77/0.18	3.79/0.14	3.71/0.06
64	2	2.38	2.94/0.00	4.33/0.00	2.41/0.00	2.86/0.00	3.40/0.00
128	12	1.62	6.02/0.58	7.25/0.66	7.20/0.64	6.94/0.63	6.58/0.65
128	8	1.65	5.97/0.57	7.23/0.66	7.10/0.62	6.87/0.62	6.16/0.57
128	4	1.78	5.70/0.52	6.91/0.58	6.60/0.56	6.00/0.49	5.53/0.44
128	2	1.92	4.90/0.37	6.43/0.48	4.75/0.31	5.23/0.37	4.89/0.31
256	12	1.34	6.66/0.71	7.80/0.79	8.38/0.79	7.68/0.75	7.18/0.78
256	8	1.38	6.54/0.68	7.72/0.77	8.02/0.75	7.92/0.78	7.23/0.79
256	4	1.47	6.32/0.64	7.64/0.75	7.76/0.71	8.07/0.81	7.18/0.78
768	2	1.31	6.68/0.71	8.01/0.83	8.42/0.80	7.97/0.79	7.34/0.81
768	1	1.54	6.00/0.58	7.35/0.68	7.25/0.64	5.81/0.46	6.44/0.63
1024	12	1.22	7.05/0.78	8.43/0.93	8.98/0.87	8.18/0.82	7.29/0.80
1024	8	1.20	7.13/0.80	8.25/0.89	8.92/0.87	8.47/0.87	7.47/0.84
1024	4	1.21	7.04/0.78	8.32/0.90	8.93/0.87	8.34/0.85	7.47/0.84
1024	2	1.27	6.68/0.71	8.22/0.88	8.52/0.81	8.04/0.80	7.24/0.79
1024	1	1.49	6.36/0.65	7.77/0.78	7.47/0.67	6.09/0.50	6.42/0.62
GPT-Neo (125M)	-	-	3.34/0.08	5.27/0.21	4.22/0.24	-	-
GPT-2-small (125M)	-	-	3.70/0.14	5.40/0.24	4.32/0.25	-	-
GPT-2-med (355M)	-	-	4.22/0.24	6.27/0.44	5.34/0.39	-	-
GPT-2-large (774M)	-	-	4.30/0.26	6.43/0.48	6.04/0.48	-	-
GPT-4	-	-	8.21/1.00	8.75/1.00	9.93/1.00	9.31/1.00	8.26/1.00

High-quality data facilitates the emergent abilities in SLMs.

Outline

- Introduction
- *Enhancement Strategy for SLMs
- *On-device SLMs and Applications
- *SLMs for LLMs
- Future Directions

On-device SLMs and Applications

On-device SLMs

Applications

Existing Generic Sub-billion SLMs

MobiLlama [6] and **MobileLLM** [7] are representative sub-billion SLMs. Why sub-billion SLMs:

- Memory constraints: An App in iPhone 15 (6GB RAM) and Google Pixel 8 Pro (12GB) should use less than 10% of RAM.

[6] Omkar et al., MobiLlama: Towards Accurate and Lightweight Fully Transparent GPT

[7] Liu et al., MobileLLM: Optimizing Sub-billion Parameter Language Models for On-Device Use Cases

Existing Generic Sub-billion SLMs

MobiLlama [6] and **MobileLLM** [7] are representative sub-billion SLMs. Why sub-billion SLMs:

- Memory constraints: An App in iPhone 15 (6GB RAM) and Google Pixel 8 Pro (12GB) should use less than 10% of RAM.
- Energy efficiency: Suppose using a 50kJ iPhone battery, at 0.1J/token per billion, and a 10 tokens/s decoding, a 7B model lasts 2 hours, while a 350M model supports a full day.

[6] Omkar et al., MobiLlama: Towards Accurate and Lightweight Fully Transparent GPT

[7] Liu et al., MobileLLM: Optimizing Sub-billion Parameter Language Models for On-Device Use Cases

Existing Generic Sub-billion SLMs

MobiLlama [6] and **MobileLLM** [7] are representative sub-billion SLMs. Why sub-billion SLMs:

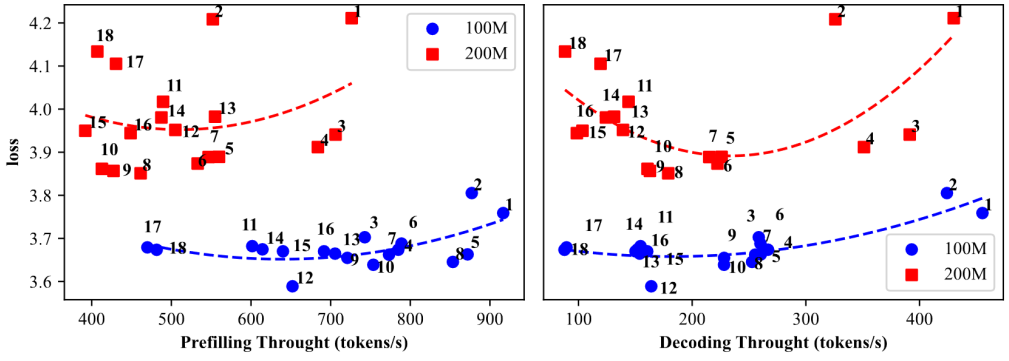
- Memory constraints: An App in iPhone 15 (6GB RAM) and Google Pixel 8 Pro (12GB) should use less than 10% of RAM.
- Energy efficiency: Suppose using a 50kJ iPhone battery, at 0.1J/token per billion, and a 10 tokens/s decoding, a 7B model lasts 2 hours, while a 350M model supports a full day.
- Decoding speed: Increases from 3-6 tokens/s for 7B models to 50 tokens/s for 125M models.

[6] Omkar et al., MobiLlama: Towards Accurate and Lightweight Fully Transparent GPT

[7] Liu et al., MobileLLM: Optimizing Sub-billion Parameter Language Models for On-Device Use Cases

Existing Generic SLMs - PhoneLM (0.5B/1.5B) [8]

A principle for SLM selection: *SLM shall adapt to the target device hardware.*



Runtime speed is more sensitive to the SLM architecture than the loss.

hidden	intermediate	layers	prefilling (tokens/s)	decoding (tokens/s)
2048	12288	16	70.75	55.12
2560	7680	18	64.98	60.60
2560	**6816**	**19**	**81.47**	**58.08**
2048	10240	19	68.52	54.48
1792	10752	21	65.42	50.18
2048	8192	22	67.10	54.04
1792	8960	25	63.29	48.63

Pre-test results for runtime speed.

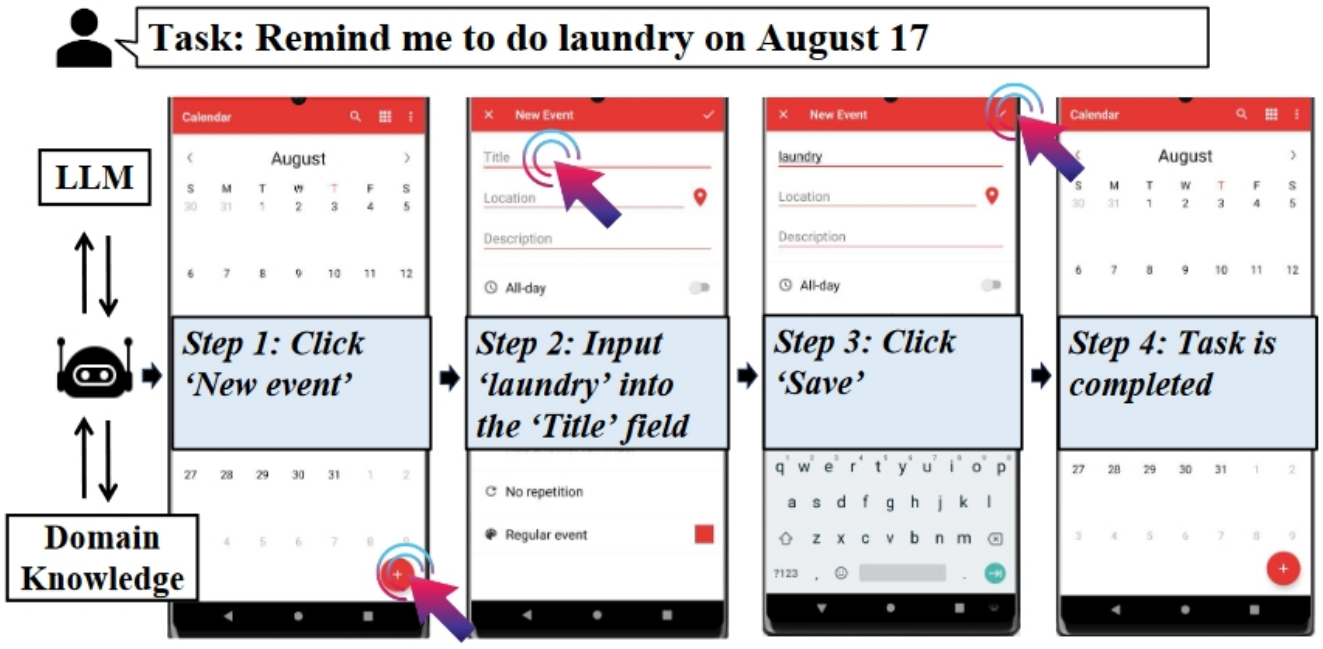
[8] Yi et al., PhoneLM:an Efficient and Capable Small Language Model Family through Principled Pre-training

On-device Applications

Generic SLMs

Applications

Why Mobile Control and Challenges



- Motivation: Hands-free.
- Challenge: Relying on the developers' API function design.
- Method: LMs utilize GUIs.

Figure credit: Carreira et al., Revolutionizing Mobile Interaction: Enabling a 3 Billion Parameter GPT LLM on Mobile.

Outline

- Introduction
- *Enhancement Strategy for SLMs
- *On-device SLMs and Applications
- *SLMs for LLMs
- Future Directions

SLMs for LLMs

Aspects SLMs help LLM:

- LLM Fine-tuning: Proxy of fine-tuning LLMs.
- LLM Evaluation: Evaluator.
- LLM Safety: lightweight safeguard.
- LLM Application: Knowledge injection.

Outline

- Introduction
- *Enhancement Strategy for SLMs
- *On-device SLMs and Applications
- *SLMs for LLMs
- Future Directions

Future Directions

- High-Quality Data Generation from LLMs: Data quality is crucial for fine-tuning.
- Personalized On-Device Models: LoRA enables tailored, lightweight parameter changes to meet personalized needs.
- Efficient Enhancement of LLMs via Proxy SLMs: Updating LLMs is costly; using SLMs for operations like optimization, knowledge integration, and data selection can serve as cost-effective proxies.
- Cloud-Edge Synergy: Edge SLMs process private data while cloud LLMs manage general data.

References

Zhenyan Lu, Xiang Li, Dongqi Cai, Rongjie Yi, Fangming Liu, Xiwen Zhang, Nicholas D Lane, and Mengwei Xu. Small language models: Survey, measurements, and insights. *arXiv preprint arXiv:2409.15790*, 2024.

Chien Van Nguyen, Xuan Shen, Ryan Aponte, Yu Xia, Samyadeep Basu, Zhengmian Hu, Jian Chen, Mihir Parmar, Sasidhar Kunapuli, Joe Barrow, et al. A survey of small language models. *arXiv preprint arXiv:2410.20011*, 2024.

Lihu Chen and Gaël Varoquaux. What is the role of small models in the llm era: A survey. *arXiv preprint arXiv:2409.06857*, 2024.

Zechun Liu, Changsheng Zhao, Forrest Iandola, Chen Lai, Yuandong Tian, Igor Fedorov, Yunyang Xiong, Ernie Chang, Yangyang Shi, Raghuraman Krishnamoorthi, Liangzhen Lai, and Vikas Chandra. MobileLLM: Optimizing sub-billion parameter language models for on-device use cases. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=EIGbXbxcUQ.

Yao Fu, Hao Peng, Litu Ou, Ashish Sabharwal, and Tushar Khot. Specializing smaller language models towards multi-step reasoning. In *International Conference on Machine Learning*, pages 10421–10430. PMLR, 2023.

Emil Emilsson. Emil is a cool guy. *Nature*, 627(9842):1–39023, 2016.

Omkar Thawakar, Ashmal Vayani, Salman Khan, Hisham Cholakal, Rao M Anwer, Michael Felsberg, Tim Baldwin, Eric P Xing, and Fahad Shahbaz Khan. Mobillama: Towards accurate and lightweight fully transparent gpt. *arXiv preprint arXiv:2402.16840*, 2024.

Thanks