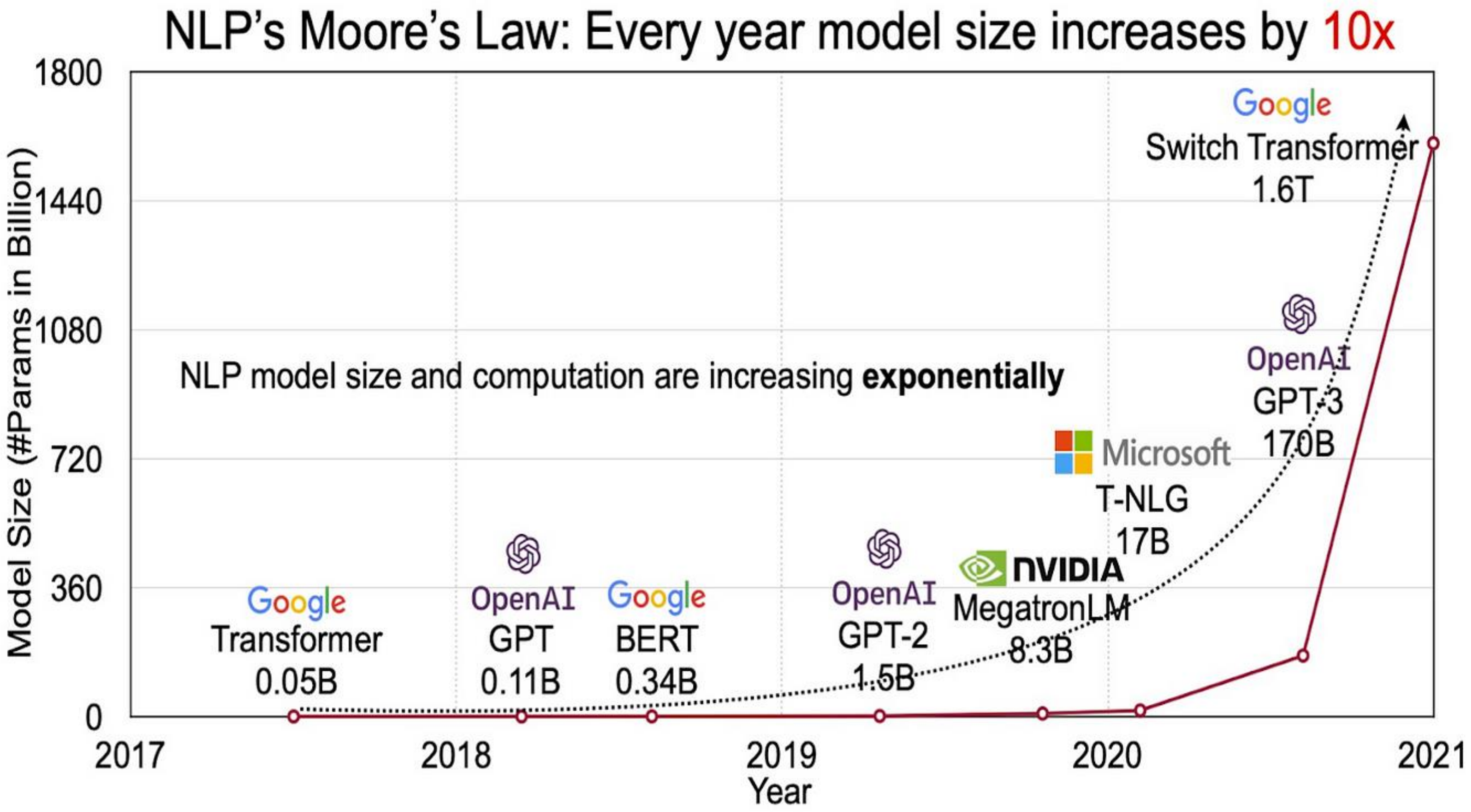
北京郵電大學
Beijing University of Posts and Telecommunications

A Few Thoughts on Small Language Models

Mengwei Xu
BUPT

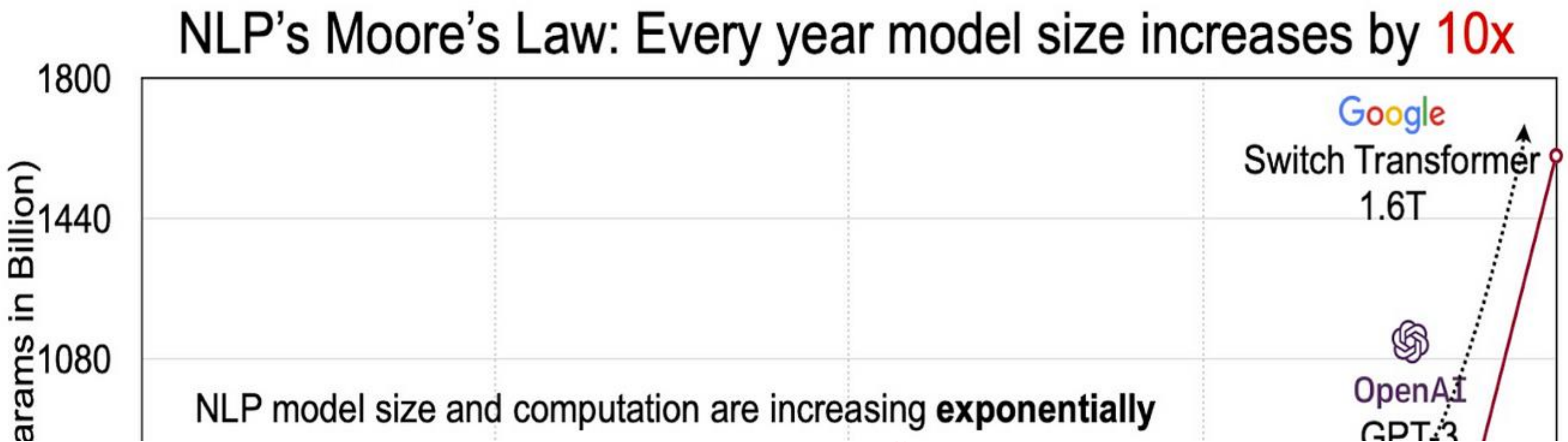
The language model evolution is diverging

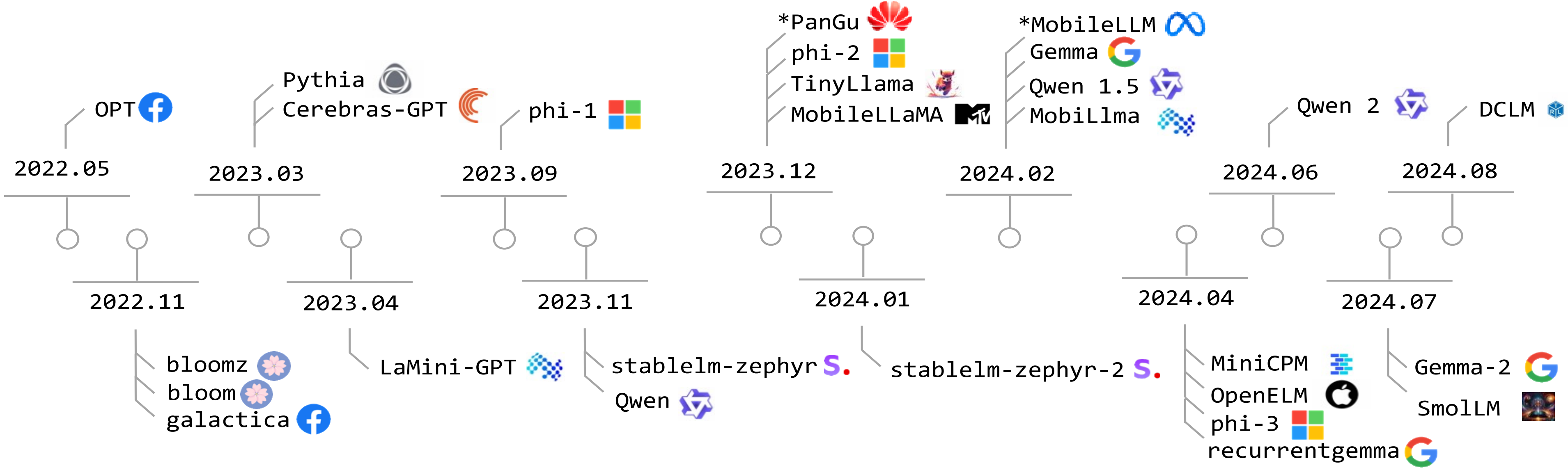
Large Language Model (LLM)



The language model evolution is diverging

Large Language Model (LLM)





2022.05 OPT

2022.11 bloomz, bloom, galactica

2023.03 Pythia, Cerebras-GPT

2023.04 LaMini-GPT

2023.09 phi-1

2023.11 stablelm-zephyr, Qwen

2023.12 *PanGu, phi-2, TinyLlama, MobileLLaMA

2024.01 stablelm-zephyr-2

2024.02 *MobileLLM, Gemma, Qwen 1.5, MobiLlma

2024.04 MiniCPM, OpenELM, phi-3, recurrentgemma

2024.06 Qwen 2

2024.07 Gemma-2, SmolLM

2024.08 DCLM

Small Language Model (SLM)

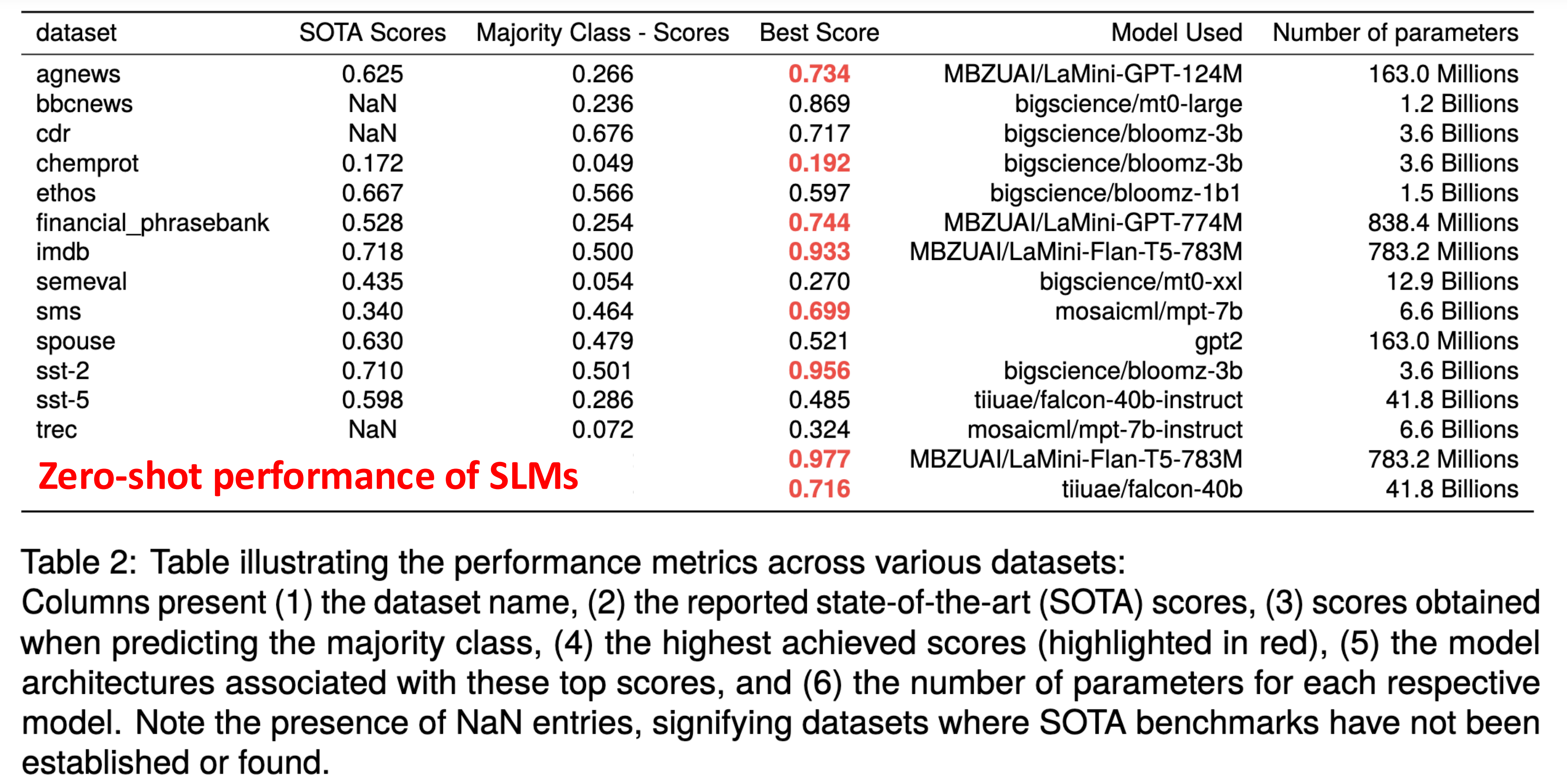
1) Can SLM really help..?
2) What makes a good SLM?
3) Device + SLM: heading to where?

Q1

Can SLM really help..?

Goal: to make devices really intelligent

SLM can solve (most) NLP tasks



dataset	SOTA Scores	Majority Class - Scores	Best Score	Model Used	Number of parameters
agnews	0.625	0.266	0.734	MBZUAI/LaMini-GPT-124M	163.0 Millions
bbcnews	NaN	0.236	0.869	bigscience/mt0-large	1.2 Billions
cdr	NaN	0.676	0.717	bigscience/bloomz-3b	3.6 Billions
chemprot	0.172	0.049	0.192	bigscience/bloomz-3b	3.6 Billions
ethos	0.667	0.566	0.597	bigscience/bloomz-1b1	1.5 Billions
financial_phrasebank	0.528	0.254	0.744	MBZUAI/LaMini-GPT-774M	838.4 Millions
imdb	0.718	0.500	0.933	MBZUAI/LaMini-Flan-T5-783M	783.2 Millions
semeval	0.435	0.054	0.270	bigscience/mt0-xxl	12.9 Billions
sms	0.340	0.464	0.699	mosaicml/mpt-7b	6.6 Billions
spouse	0.630	0.479	0.521	gpt2	163.0 Millions
sst-2	0.710	0.501	0.956	bigscience/bloomz-3b	3.6 Billions
sst-5	0.598	0.286	0.485	tiiuae/falcon-40b-instruct	41.8 Billions
trec	NaN	0.072	0.324	mosaicml/mpt-7b-instruct	6.6 Billions
			0.977	MBZUAI/LaMini-Flan-T5-783M	783.2 Millions
			0.716	tiiuae/falcon-40b	41.8 Billions

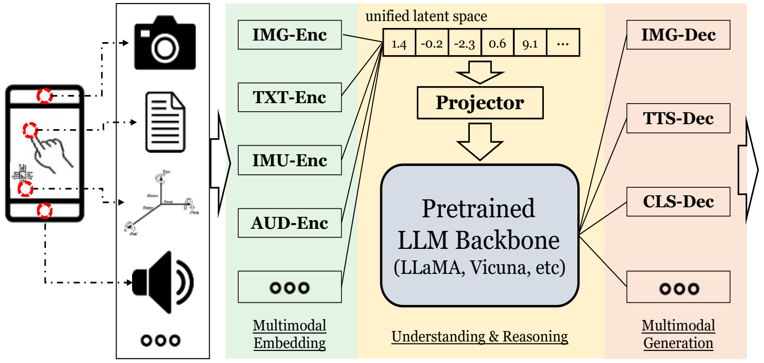
Zero-shot performance of SLMs

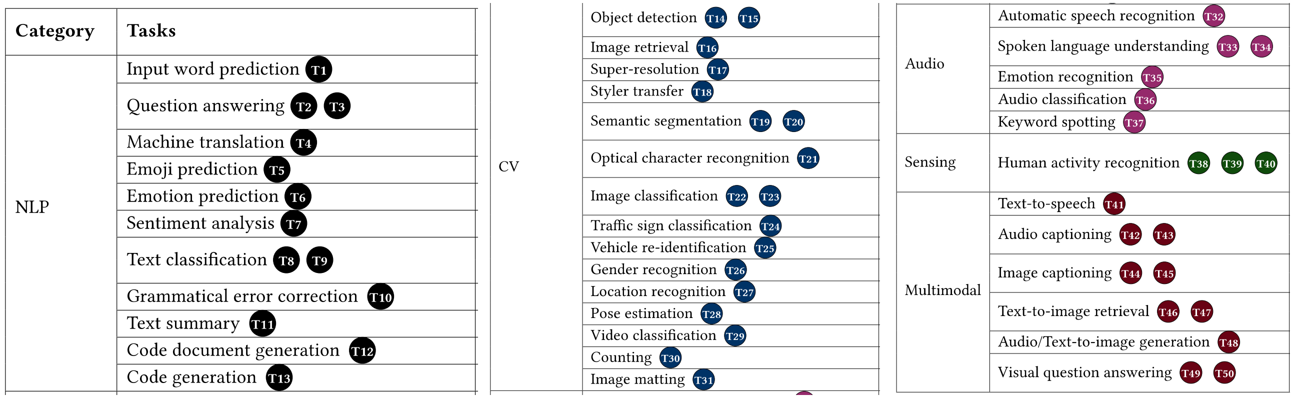
Table 2: Table illustrating the performance metrics across various datasets: Columns present (1) the dataset name, (2) the reported state-of-the-art (SOTA) scores, (3) scores obtained when predicting the majority class, (4) the highest achieved scores (highlighted in red), (5) the model architectures associated with these top scores, and (6) the number of parameters for each respective model. Note the presence of NaN entries, signifying datasets where SOTA benchmarks have not been established or found.

[ACL'24] Small Language Models are Good Too: An Empirical Study of Zero-Shot Classification. Pierre Lepagnol, et al.

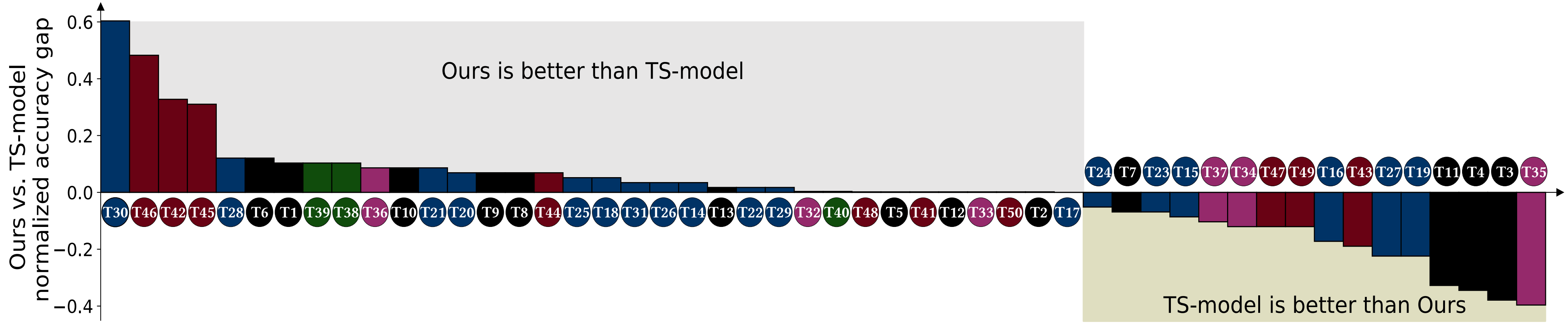
Mobile foundation model

- We trained an on-device foundation model that outperforms traditional handcrafted DNNs in 50 multimodal tasks.



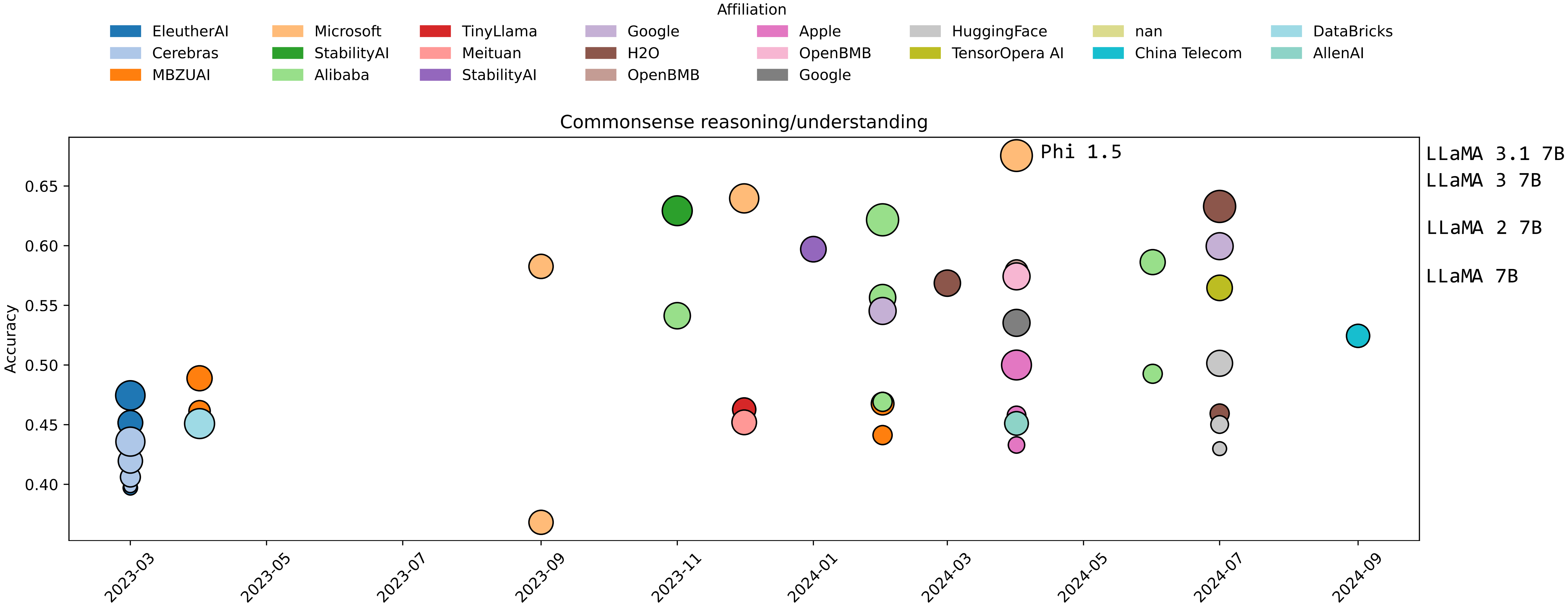


Category	Tasks
NLP	Input word prediction T1
	Question answering T2 T3
	Machine translation T4
	Emoji prediction T5
	Emotion prediction T6
	Sentiment analysis T7
	Text classification T8 T9
	Grammatical error correction T10
	Text summary T11
	Code document generation T12
	Code generation T13
CV	Object detection T14 T15
	Image retrieval T16
	Super-resolution T17
	Styler transfer T18
	Semantic segmentation T19 T20
	Optical character recongnition T21
	Image classification T22 T23
	Traffic sign classification T24
	Vehicle re-identification T25
	Gender recognition T26
	Location recognition T27
	Pose estimation T28
	Video classification T29
	Counting T30
	Image matting T31
Audio	Automatic speech recognition T32
	Spoken language understanding T33 T34
	Emotion recognition T35
	Audio classification T36
	Keyword spotting T37
Sensing	Human activity recognition T38 T39 T40
Multimodal	Text-to-speech T41
	Audio captioning T42 T43
	Image captioning T44 T45
	Text-to-image retrieval T46 T47
	Audio/Text-to-image generation T48
	Visual question answering T49 T50



[MobiCom’24] Mobile Foundation Model as Firmware. Jinliang Yuan, et al.

SLM is evolving fast



Q1

Can SLM really help..?

Goal: to make devices really intelligent

Insight

SLM can solve most (if not all) mobile AI tasks, thanks to the combination of efficient architecture (decoder-only Transformer) + training method (next-token pred.)

Q2

What makes a good SLM?

Goal: high accuracy, low cost

An overall guideline

- **Model size -> memory**
- Model architecture -> speed
- Data quality/quantity -> capability

1. Offloading with hierarchical storage on devices alone can hardly help
   - Gap between DRAM and disk is too large
2. Combined with sparsity (MoE, activation sparsity, etc), is probably the way to go[1]
   - But can MoE models scale efficiently?[2]

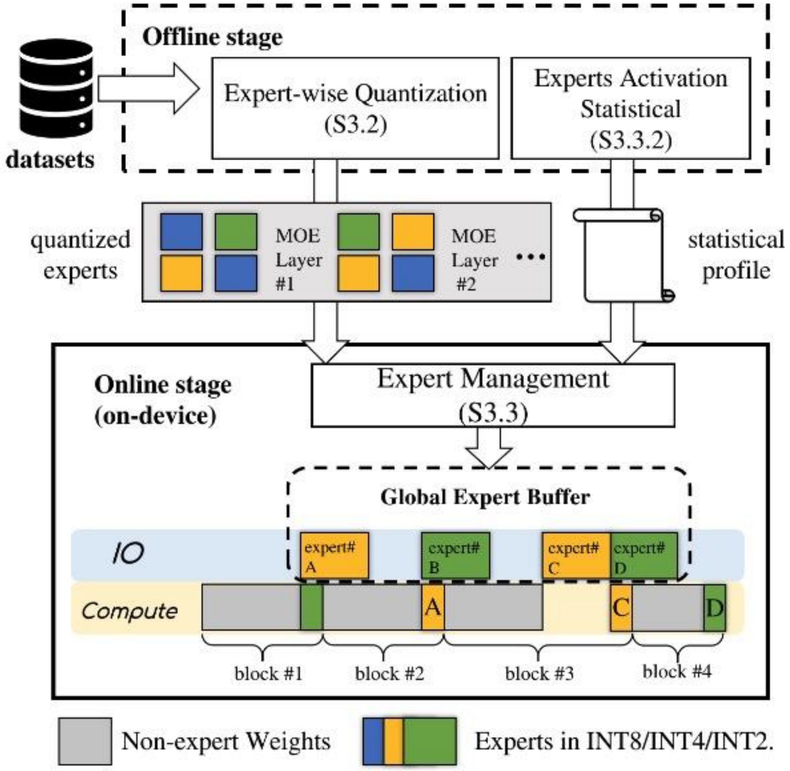


Figure 5: System architecture of EdgeMoE and workflow.

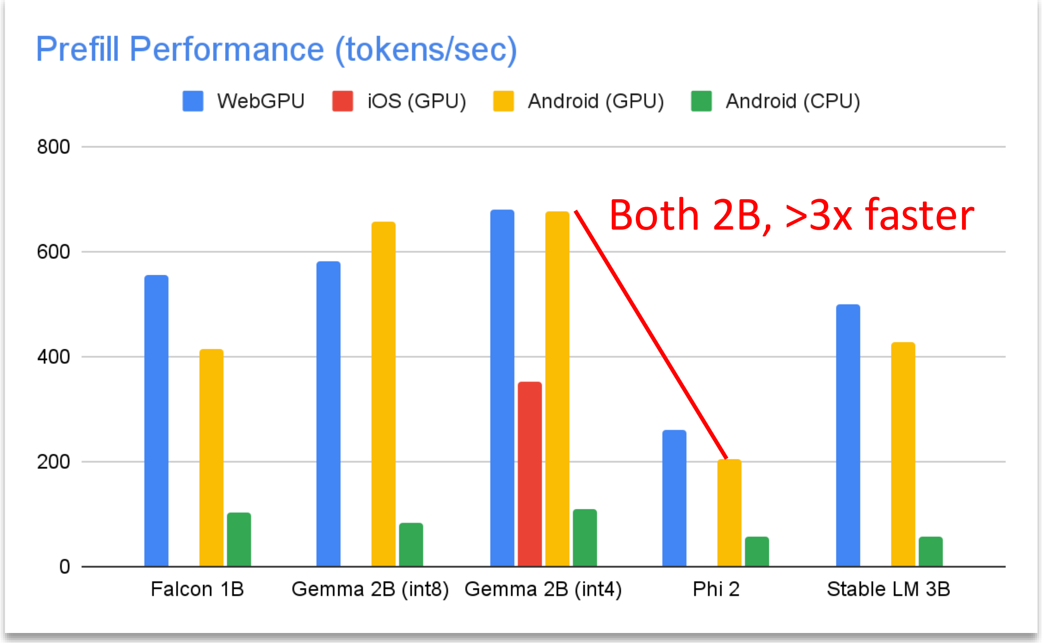
[1] EdgeMoE: Fast On-Device Inference of MoE-based Large Language Models, Rongjie Yi, et al

[2] Scaling Laws for Fine-grained Mixture of Experts, Jan Ludziejewski, et al

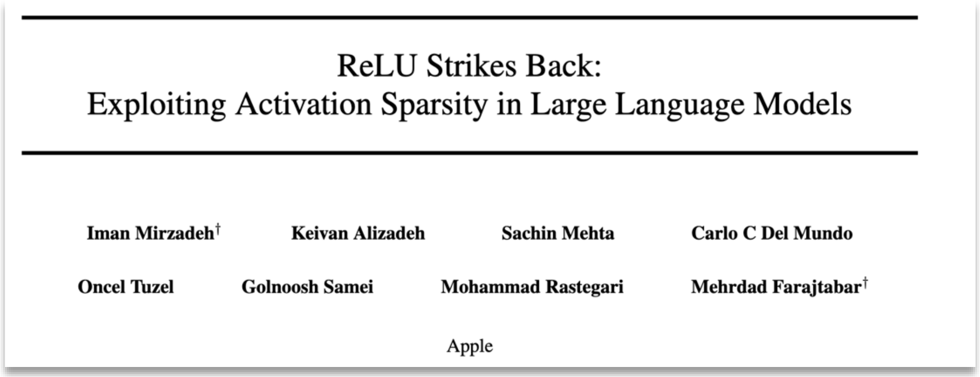
An overall guideline

- Model size -> memory
- **Model architecture -> speed**
- Data quality/quantity -> capability

- The configuration matters (width vs height, # of heads, hidden size, etc)
- The quantization schema matters, especially on hardware accelerators



Both 2B, >3x faster



ReLU Strikes Back:
Exploiting Activation Sparsity in Large Language Models

Iman Mirzadeh[†] Keivan Alizadeh Sachin Mehta Carlo C Del Mundo
Oncel Tuzel Golnoosh Samei Mohammad Rastegari Mehrdad Farajtabar[†]
Apple

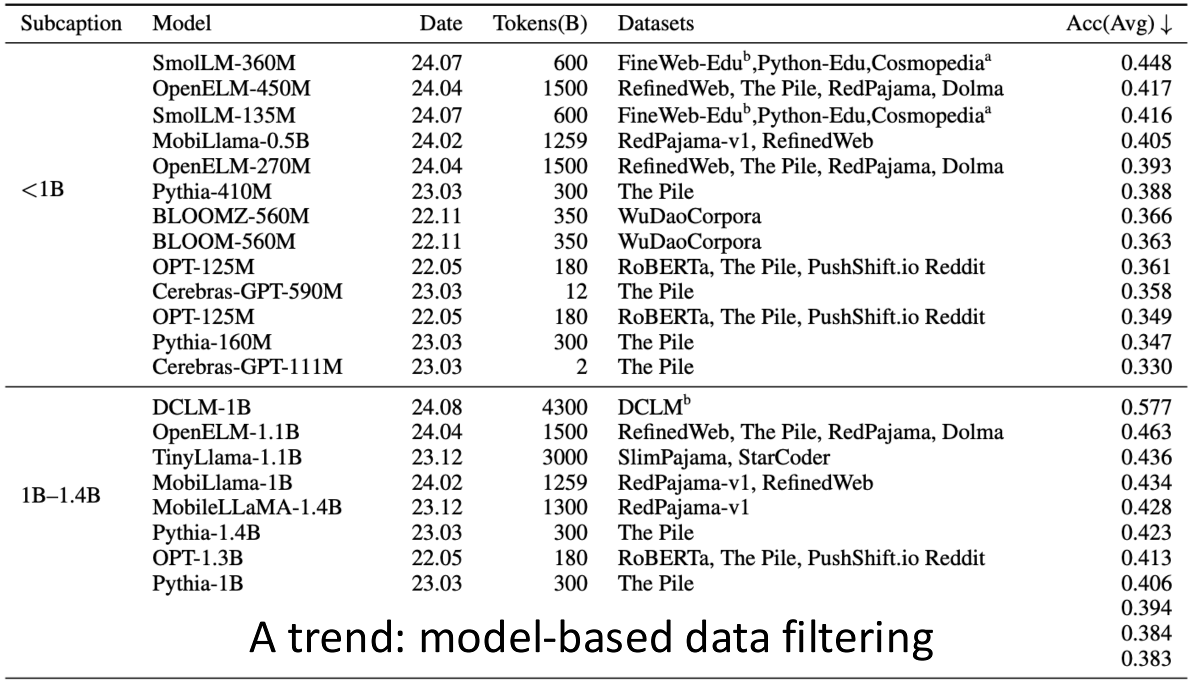
Empowering 1000 tokens/second on-device LLM prefilling with mllm-NPU

Daliang Xu[♦], Hao Zhang[◇], Liming Yang[♦], Ruiqi Liu[♦], Gang Huang[♦], Mengwei Xu[◇*], Xuanzhe Liu[♦]
[♦]Peking University, [◇]Beijing University of Posts and Telecommunications
Code and demo: https://github.com/UbiquitousLearning/mllm

https://developers.googleblog.com/en/large-language-models-on-device-with-mediapipe-and-tensorflow-lite/

An overall guideline

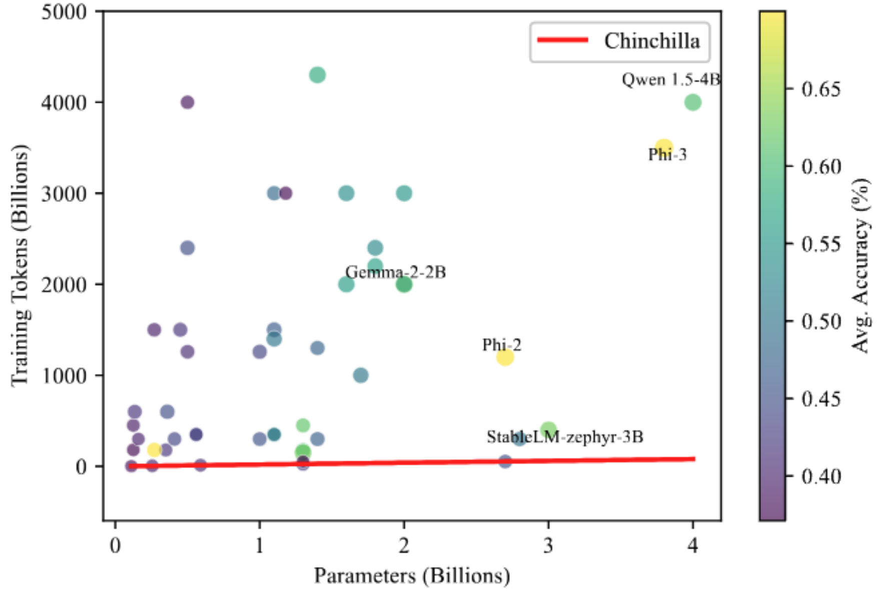
- Model size -> memory
- Model architecture -> speed
- **Data quality/quantity -> capability**



Subcaption	Model	Date	Tokens(B)	Datasets	Acc(Avg) ↓
<1B	SmolLM-360M	24.07	600	FineWeb-Edu[b],Python-Edu,Cosmopedia[a]	0.448
	OpenELM-450M	24.04	1500	RefinedWeb, The Pile, RedPajama, Dolma	0.417
	SmolLM-135M	24.07	600	FineWeb-Edu[b],Python-Edu,Cosmopedia[a]	0.416
	MobiLlama-0.5B	24.02	1259	RedPajama-v1, RefinedWeb	0.405
	OpenELM-270M	24.04	1500	RefinedWeb, The Pile, RedPajama, Dolma	0.393
	Pythia-410M	23.03	300	The Pile	0.388
	BLOOMZ-560M	22.11	350	WuDaoCorpora	0.366
	BLOOM-560M	22.11	350	WuDaoCorpora	0.363
	OPT-125M	22.05	180	RoBERTa, The Pile, PushShift.io Reddit	0.361
	Cerebras-GPT-590M	23.03	12	The Pile	0.358
	OPT-125M	22.05	180	RoBERTa, The Pile, PushShift.io Reddit	0.349
	Pythia-160M	23.03	300	The Pile	0.347
	Cerebras-GPT-111M	23.03	2	The Pile	0.330
1B–1.4B	DCLM-1B	24.08	4300	DCLM[b]	0.577
	OpenELM-1.1B	24.04	1500	RefinedWeb, The Pile, RedPajama, Dolma	0.463
	TinyLlama-1.1B	23.12	3000	SlimPajama, StarCoder	0.436
	MobiLlama-1B	24.02	1259	RedPajama-v1, RefinedWeb	0.434
	MobileLLaMA-1.4B	23.12	1300	RedPajama-v1	0.428
	Pythia-1.4B	23.03	300	The Pile	0.423
	OPT-1.3B	22.05	180	RoBERTa, The Pile, PushShift.io Reddit	0.413
	Pythia-1B	23.03	300	The Pile	0.406
					0.394
					0.384
					0.383

A trend: model-based data filtering

A trend: “over-training”



(a) The relationship between Training Tokens and Parameters.

Q2

What makes a good SLM?

Goal: high accuracy, low cost

Insight

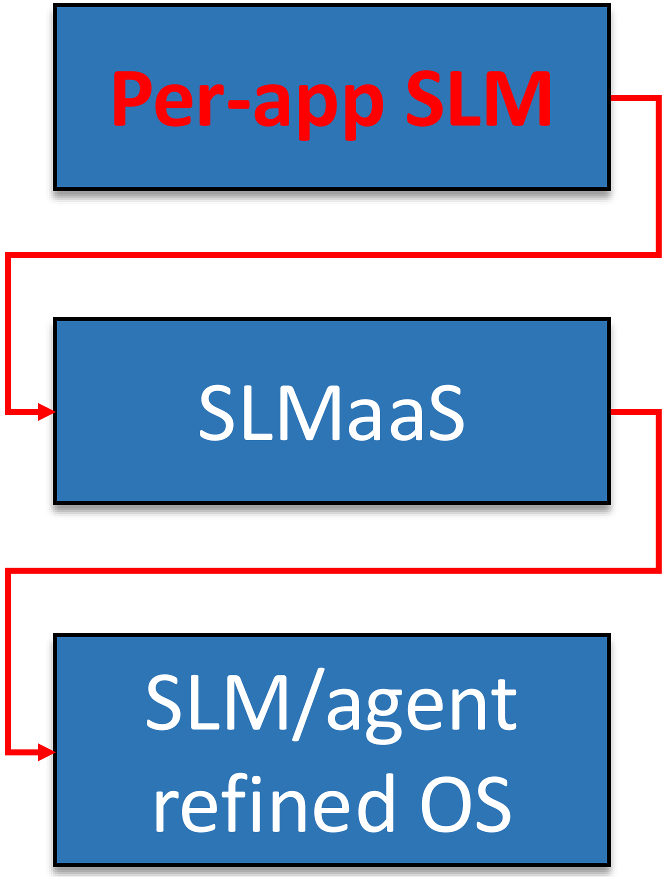
Search for a hardware-friendly architecture, trained with high-quality/quantity data

Q3

Device + SLM: heading to where?

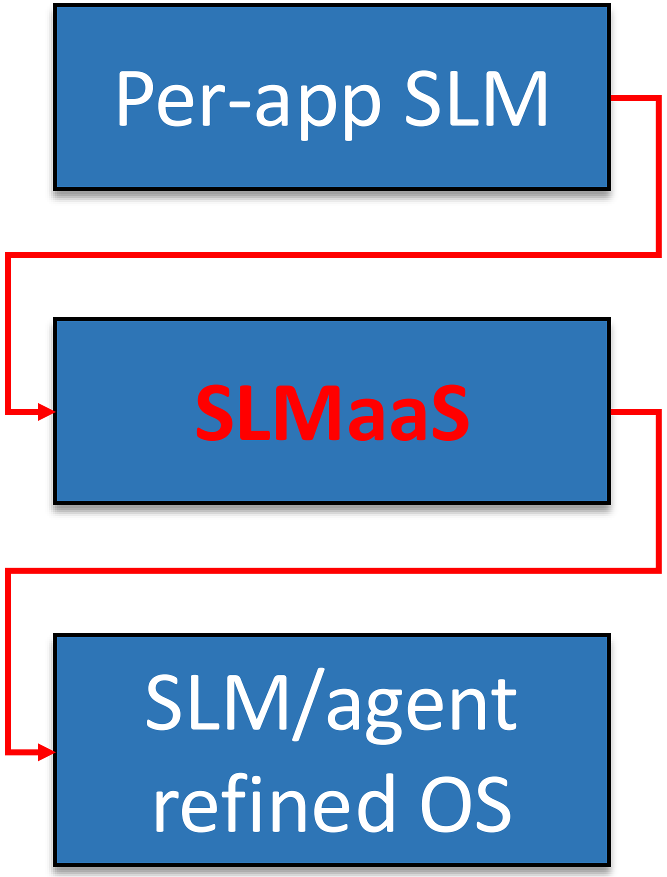
Goal: fully unleash the power of SLM and device for agents

How fundamentally will SLM (w/ cloud LLM) subvert the devices system (and OS)



- Limited scalability (memory constraint), Huge deployment cost
- Difficult to enjoy hardware acceleration
- OS-transparent – no cross-app batching/caching/scheduling
- AI not democratized ☺

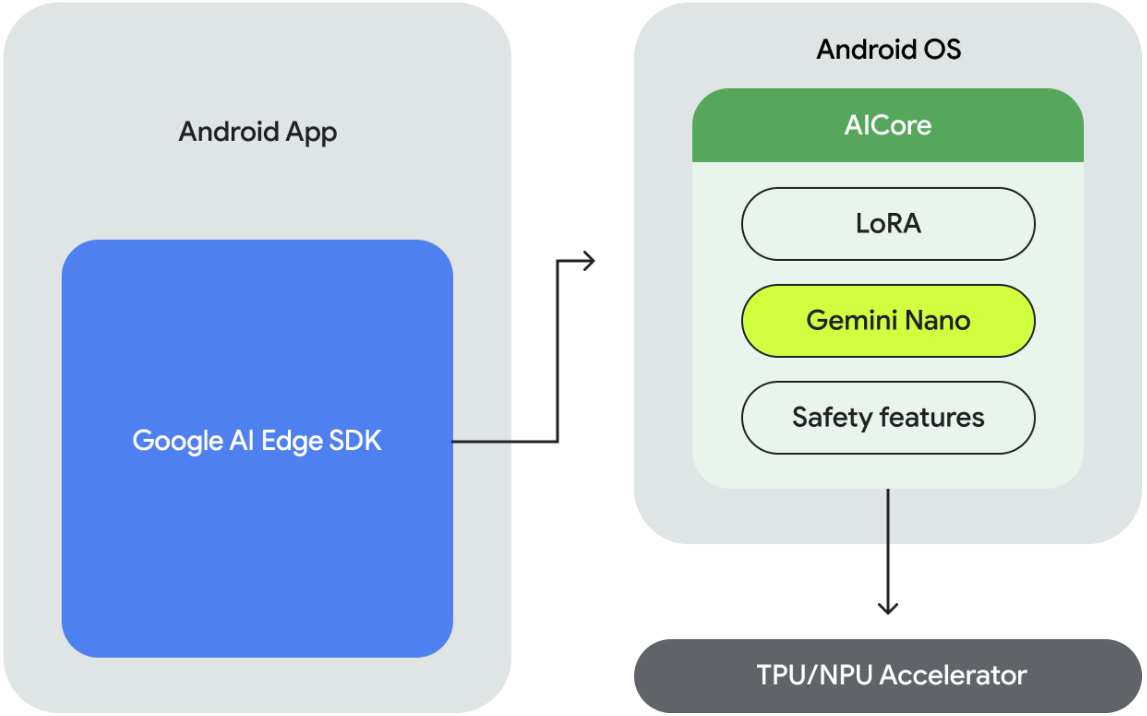
How fundamentally will SLM (w/ cloud LLM) subvert the devices system (and OS)



Per-app SLM

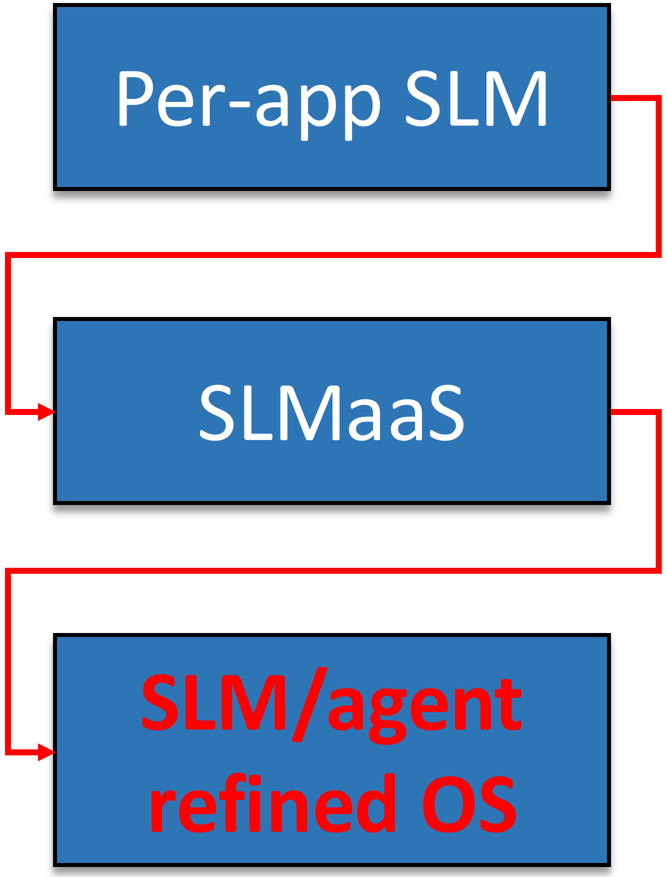
SLMaaS

SLM/agent refined OS



[1] https://developer.android.com/ml/aicore

How fundamentally will SLM (w/ cloud LLM) subvert the devices system (and OS)



Per-app SLM

SLMaaS

SLM/agent refined OS

Let’s go beyond this!

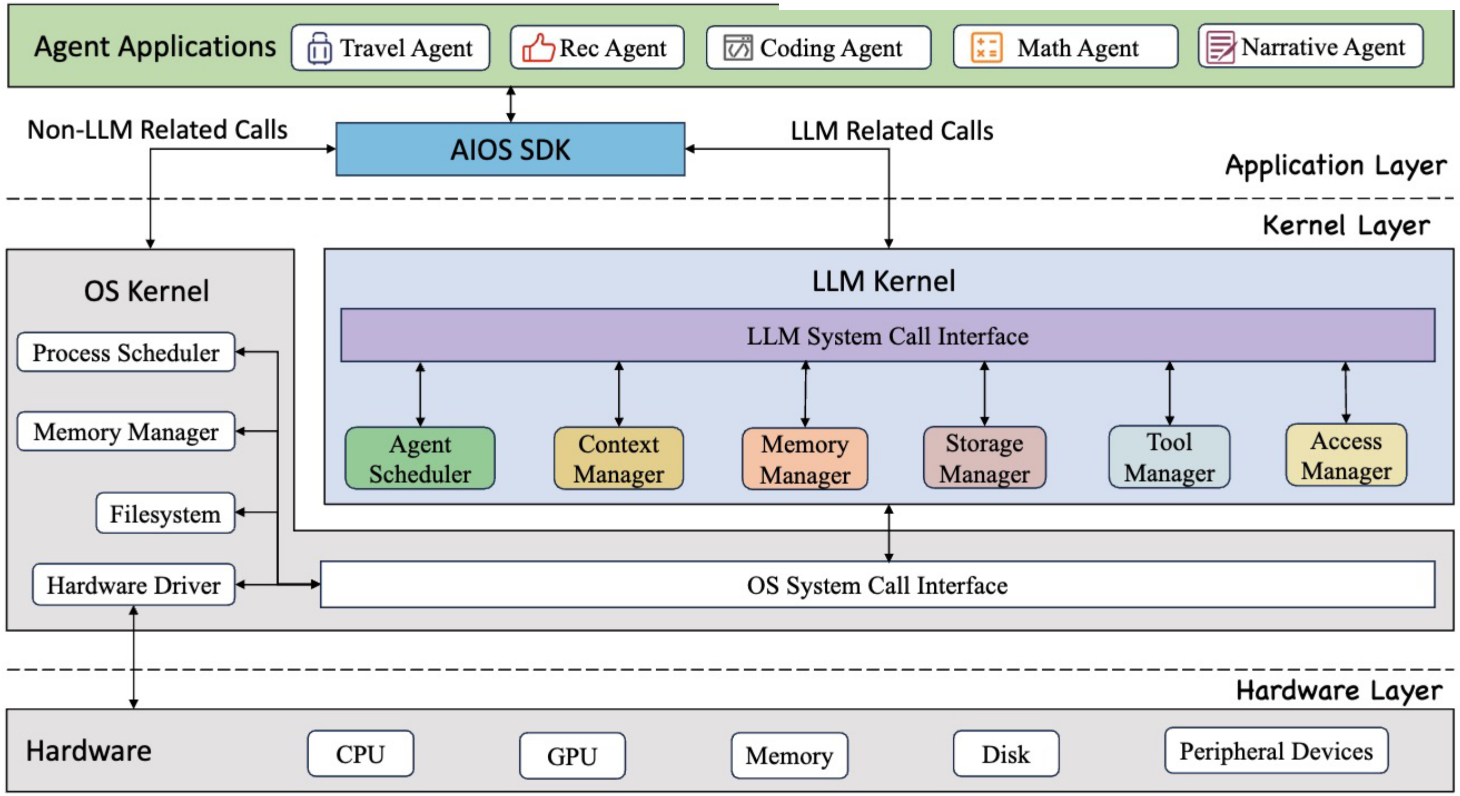
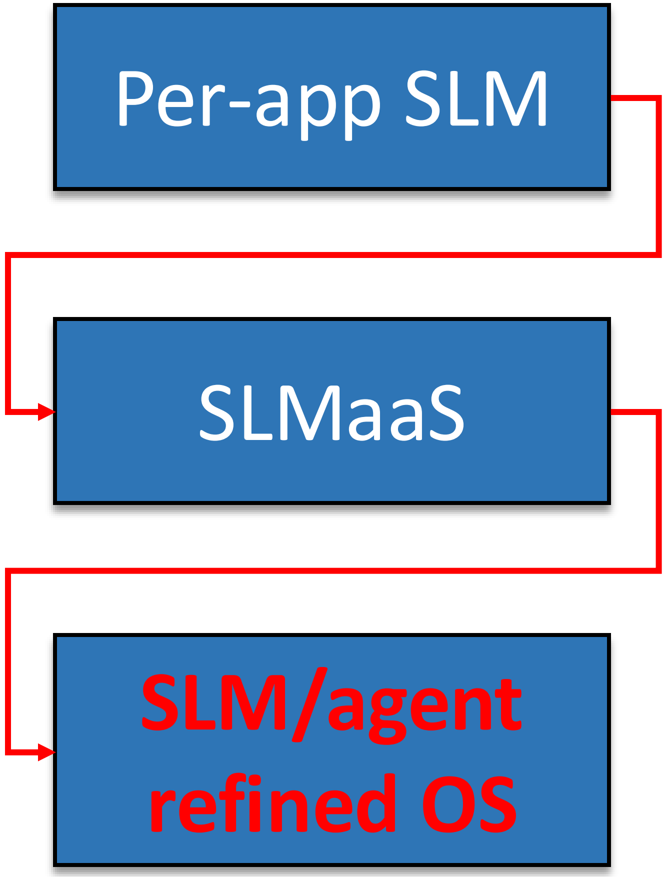


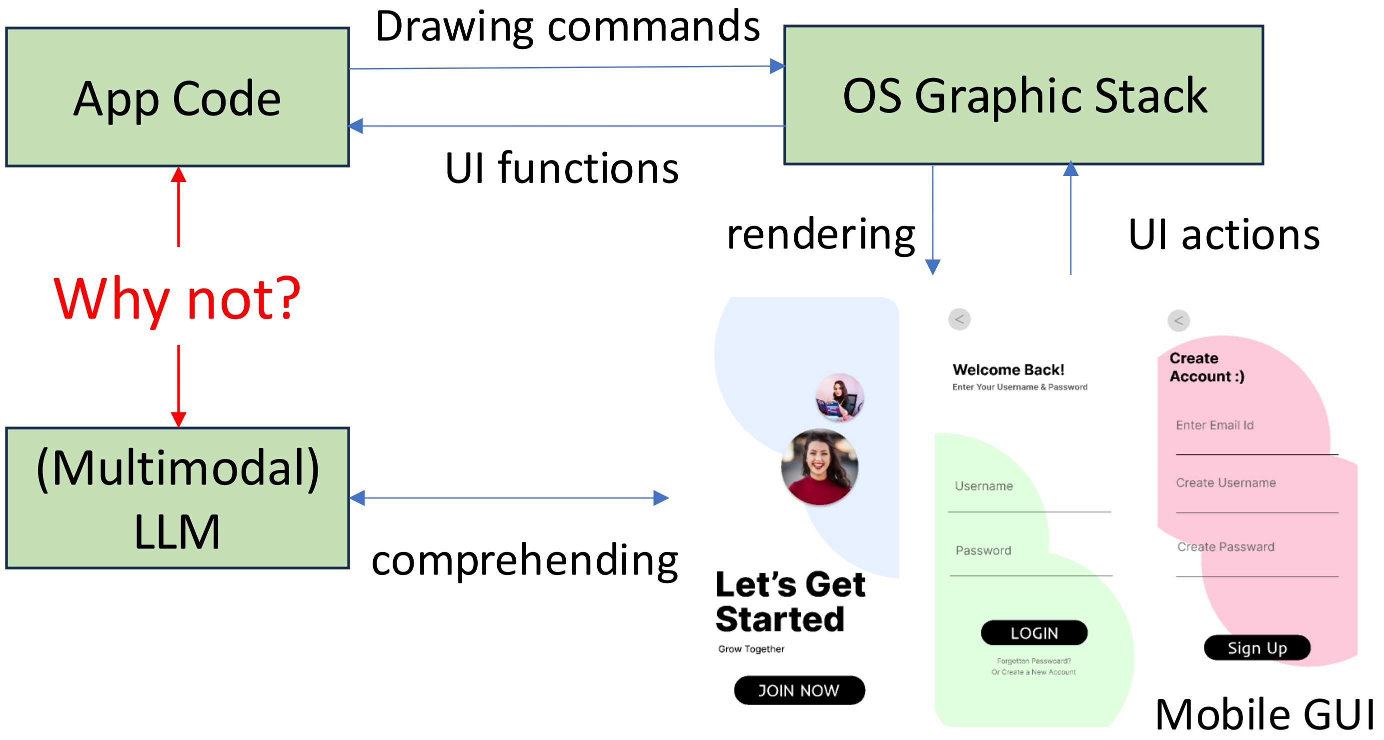
Figure 2: An overview of the AIOS architecture.

[arXiv’24] AIOS: LLM Agent Operating System, Kai Mei, et al.

How fundamentally will SLM (w/ cloud LLM) subvert the devices system (and OS)



Taking GUI Agent as an example



Q3

Device + SLM: heading to where?

Goal: fully unleash the power of SLM and device for agents

Insight

Deep integration of SLM with smart devices software stack; SLM and agents might redefine OSes.