
SMALL LANGUAGE MODELS: SURVEY, MEASUREMENTS, AND INSIGHTS

Zhenyan Lu^{♣◊†}, Xiang Li^{♣†}, Dongqi Cai[♣], Rongjie Yi[♣], Fangming Liu[◊], Xiwen Zhang[♥],
Nicholas D. Lane[♡], Mengwei Xu[♣]

[♣]Beijing University of Posts and Telecommunications (BUPT)

[◊]Peng Cheng Laboratory

[♥]Helixon Research

[♡]University of Cambridge

Website: https://github.com/UbiquitousLearning/SLM_Survey

ABSTRACT

Small language models (SLMs), despite their widespread adoption in modern smart devices, have received significantly less academic attention compared to their large language model (LLM) counterparts, which are predominantly deployed in data centers and cloud environments. While researchers continue to improve the capabilities of LLMs in the pursuit of artificial general intelligence, SLM research aims to make machine intelligence more accessible, affordable, and efficient for everyday tasks. Focusing on transformer-based, decoder-only language models with 100M–5B parameters, we survey 70 state-of-the-art open-source SLMs, analyzing their technical innovations across three axes: architectures, training datasets, and training algorithms. In addition, we evaluate their capabilities in various domains, including commonsense reasoning, mathematics, in-context learning, and long context. To gain further insight into their on-device runtime costs, we benchmark their inference latency and memory footprints. Through in-depth analysis of our benchmarking data, we offer valuable insights to advance research in this field.

Keywords Small Language Model · Edge Intelligence · On-device LLM

1 Introduction

The evolution of language models is diverging. On one hand, in the pursuit of artificial general intelligence (AGI) and following the scaling law, increasingly large language models (LLM) have been born in datacenters that host hundreds of thousands of GPUs [40, 94]. The aim of this path is to demonstrate that machines can solve the most challenging language tasks, with the ultimate goal of advancing human civilization by pushing the boundaries of science and technology. On the other hand, there is a growing focus on **small language models (SLM)**, designed for resource-efficient deployment on personal devices such as desktops, smartphones, and even wearables. The vision behind SLMs is to democratize access to machine intelligence, making it both accessible and affordable to people everywhere. This approach seeks to make intelligence ubiquitous and practical, available to anyone, anywhere, at any time – much like the human brain, which everyone possesses.

Both LLM and SLM are important in reshaping our daily lives, yet the latter receives significantly less attention in academia. There has been very limited literature exploring SLM capabilities [44, 74, 108] or their runtime cost on devices [50, 43, 92], often with limited scale or depth. In the real world, however, SLMs have already been integrated into commercial off-the-shelf (COTS) devices on a massive scale [99, 25]. For instance, the latest Google/Samsung smartphones have built-in LLM services (Gemini Nano), allowing third-party mobile apps to leverage LLM capa-

[†]Zhenyan Lu and Xiang Li contributed equally to this work.

bilities through prompts and LoRA modules [1]. The most recent iOS system on iPhones and iPads also includes an on-device foundation model, deeply integrated with the operating system for better performance and privacy [2]. Beyond resource-constrained scenarios, SLMs also excel superior performance in certain domain-specific tasks [46].

This work presents the first comprehensive survey of SLMs, thoroughly discussing their capabilities, runtime costs, and innovations in recent years. The scope of this survey is limited to those *language models with 100M–5B parameters in decoder-only transformer architecture*^{*}, which covers the range of devices from low-end IoT/wearable gadgets like smartwatches to high-end mobile devices such as smartphones and tablets. In total, we collected 57 popular SLMs and fully benchmarked their capabilities (commonsense reasoning, math, in-context learning, long-context retrieval, etc.) and on-device runtime costs (prefill and decode speed, memory footprint, energy consumption, etc.).

With the insights from comprehensive benchmarking and thorough investigation, We try to answer the following questions concerning SLMs: “What is the evolving path of SLMs?” “What datasets or training strategies are more likely to produce a highly capable SLM?” “How different SLM architecture (e.g., depth, width, atten type) and the deployment environments (quantization algorithms, hardware type, etc) impact runtime performance?” We expect the work to reveal an all-sided landscape of SLM and benefit the research community, including those working on the algorithm, model, system, and hardware levels.

In summary, we make the following contributions in this work.

- We exhaustively review the small language models released in recent years, summarize their key innovations, and benchmark their capability as well as on-device cost.
- Through such in-depth investigation, we obtain valuable insights from open-sourced SLMs, which can potentially benefit future SLM research. We also summarize a few potential research topics in SLM.
- We make all results and benchmark tools public to advance and facilitate the SLM research.

2 SLM Architecture, Datasets, and Training

2.1 Overview of SLMs

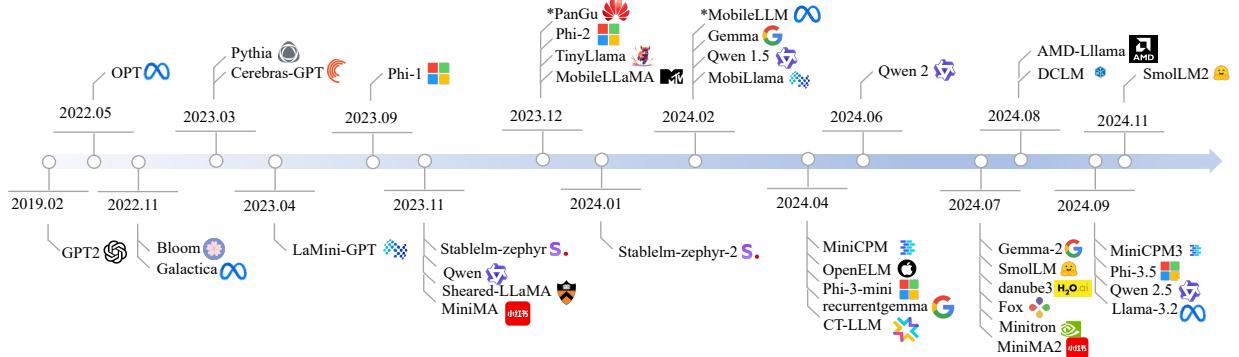


Figure 1: An overview of SLMs. * indicates the models are not open-sourced so will not be benchmarked.

As shown in Figure 1, SLMs have gained increasing attention from both the research and industrial communities. Notably, since the end of 2023, the number of SLM models has surged significantly. To understand their capability and cost, we comprehensively collect SLMs based on the following criteria: (1) We only collect models with decoder-only transformer architecture (which can be traced to GPT-2), for their superior performance and real-world deployment. Currently we do not include variants of transformers such as RWKV [71] and Mamba [34]. (2) We only collect SLMs with open weights, so that we can evaluate them. (3) The weight range of SLMs in this work is defined between 100M to 5B. (4) The survey mainly focuses on the base knowledge obtained from pre-training process, thereby we only collect the base models of SLMs, except those provided only the instruct versions (Microsoft Phi and StabilityAI StableLM). We also exclude models that are fine-tuned on other pre-trained models.

^{*}The definition of “small” could drift over time, considering that device memory is increasing over time and can host larger “small language models” in the future. We set 5B as the upper limit for the size of SLMs, since as of Seq. 2024, 7B LLMs are still mostly deployed in the cloud.

Affiliation	Model name	Size	Date	Attention	Layer number	Hidden size	Head num	Activation	Vocab. size	Open training datasets	Max context window
Meta 	OPT [28]	125M 350M 1.3B 2.7B	2022.05	MHA	12 24 24 32	768 1024 2048 2560	12 16 32 32	ReLU	50k	✓	2k
		125M 1.3B			12 24	768 2048	12 32				
	Llama-3.2 [59]	3B 1B	2024.09 2024.09	GQA GQA	28 16	3072 2048	24 32	SiLU SiLU	128k 128k		131k 131k
BigScience 	Bloom [15]	560M 1.1B	2022.11	MHA	24	1024 1536	16	GELU _{tanh}	251k	✓	2k
	Bloomz [16]	1.1B 560M	2022.11	MHA		1536 1024					
EleutherAI 	Pythia [27]	160M 410M 1B 1.4B 2.8B	2023.03	MHA	12 24 16 24 32	768 1024 2048 2048 2560	12 16 8 16 32	GELU	50k	✓	2k
Cerebras 		111M 256M 590M 1.3B 2.7B			10 14 18 24 32	768 1088 1536 2048 2560	12 17 12 16 32				
Microsoft 	Phi-1 [60]	1.3B	2023.09	MHA	24	2048	32	GELU _{tanh}	51k		2k
	Phi-1.5 [61]	1.3B	2023.09	MHA	24	2048	32	GELU _{tanh}	51k		
	Phi-2 [62]	2.7B	2023.12	MHA	32	2560	32	GELU _{tanh}	51k		2k
	Phi-3-mini* [63]	3.8B	2024.04	MHA	32	3072	32	SiLU	32k		4k
	Phi-3.5-mini*	2.7B	2024.09	MHA	32	3072	32	SiLU	32k		4k
StabilityAI 	StableLM-zephyr* [79]	3B	2023.11	MHA	32	2560	32	SiLU	50k	✓	1k
	StableLM-2-zephyr* [80]	1.6B	2024.01	MHA	24	2048	32	SiLU	100k	✓	4k
TinyLlama 	TinyLlama [105]	1.1B	2023.12	GQA	22	2048	32	SiLU	32k	✓	2k
Meituan 	MobileLLaMA [58]	1.4B	2023.12	GQA	24	2048	16	SiLU	32k	✓	2k
Alibaba 	Qwen 1 [3]	1.8B	2023.11	MHA	24	2048	16	SiLU	152k		8k
	Qwen 1.5 [4]	0.5B	2024.02	MHA	24	1024	16	SiLU	152k		32k
	Qwen 2 [5]	1.8B 4B	2024.06	MHA	24 40	2048 2560	16 20	SiLU	152k		32k
	Qwen 2.5 [6]	0.5B 1.5B 3B	2024.09	GQA	24 28 36	896 1536 2048	14 12 16	SiLU	152k		32k
MBZUAI 	MobiLlama [57]	0.5B 1B	2024.02	GQA	22	2048	32	SiLU	32k	✓	2k
	LaMini-GPT [56]	774M 1.5B	2023.04	MHA	36 48	1280 1600	20 25	GELU _{tanh}	50k		1k
Google 	Gemma [32]	2B	2024.02	MQA	18	2048	8	GELU	256k		8k
	recurrentGemma [33]	2B	2024.04	MQA	26	2560	10	GELU _{tanh}	256k		8k
	Gemma-2	2B	2024.07	GQA	26	2304	8	GELU _{tanh}	256k		8k
OpenBMB 	MiniCPM [67]	1B 2B	2024.04	GQA	52 40	1536 2304	24 36	SiLU	73k 123k		128k 131k
	MiniCPM3 [68]	4B	2024.09	MLA	62	2560	40	SiLU	73k		
Apple 	OpenELM [10]	270M 450M 1.1B 3B	2024.04	GQA	16 20 28 36	1280 1536 2048 3072	12-20 12-24 16-32 12-24	SiLU	32k	✓	2k
H2O 	danube3 [35]	0.5B 4B	2024.07	GQA	16 24	1536 3840	16 32	SiLU SiLU	32k 32k		8k 8k
TensorOpera AI 	Fox [82]	1.6B	2024.07	GQA	32	2048	16	SiLU	32k		8k
HuggingFace 	SmolLM [38]	135M 360M 1.7B	2024.07	GQA	30 32 MHA	576 960 2048	9 15 32	SiLU	49k	✓	2k
		360M 1.7B			24	2048	32				
	SmolLM2 [39]	360M 1.7B	2024.11	GQA MHA	30 24	576 2048	9 32				
Toyota 	DCLM [83]	1.4B	2024.08	MHA	24	2048	16	SiLU	50k	✓	50k
DataBricks 	Dolly-v2* [24]	3B	2023.04	MHA	32	2560	32	GELU	50k		2k
AllenAI 	OLMo [8]	1.18B	2024.04	MHA	16	2048	16	SiLU	50k	✓	50k
Princeton 	Sheared-LLaMA [72]	1.3B 2.7B	2023.11 2023.11	MHA MHA	24 32	2048 2560	16 20	SiLU SiLU	32k 32k		4096 4096
Xiaohongshu 	MiniMA [88] MiniMA2 [89]	3B 1B	2023.11 2024.07	UKN	24 18	3072 2304	24 18	SiLU SiLU	49k 49k		4096 4096
Nvidia 	Minitron [66]	4B	2024.07	GQA	32	3072	24	ReLU2	256k		4096
M.A.P. 	CT-LLM [55]	2B	2024.04	MHA	32	2048	16	SiLU	125k		4096
AMD 	AMD-Llama [9]	135M	2024.08	MHA	12	768	12	SiLU	32k		2048

Table 1: Detailed configurations of SLMs benchmarked. We mainly use the base models in experiments, with exceptions of StableLM, Phi-3/3.5 and Dolly-v2 (marked with *) that only provide the instruct version.

With such criteria, we select 70 SLMs as detailed in Table 1. Our selection encompasses a wide range of models from both industry and academia, based on factors such as model architecture, parameter size, and data availability. While all selected SLMs share similar architectures, they differ in specific hyperparameters and training datasets, with some datasets remaining closed-source. These variations lead to differing performance across tasks, as we will discuss in the following sections.

2.2 Model Architecture

While we focus on only decoder-only transformer SLMs, their specific configurations still diversify, as shown in Figure 2(a). The core of Transformer is the multi-head self-attention(MHA) mechanism and the Feed-Forward Neural Network(FFN).

Model architecture analysis. We conduct statistical analysis on the following several components of the model architecture: 1) The type of self-attention; 2) The type of feed-forward neural network; 3) The intermediate ratio of the feed-forward network; 4) The activation function of the feed-forward neural network; 5) The type of layer normalization; 6) The vocabulary size. Figure 2(a) shows the architecture of SLM and the pie chart shows the distribution of six components. Figure 2(b) shows how these distributions change over time.

1) *The type of self-attention.* The self-attention mechanism is the core of the Transformer model. In general, SLMs mainly use three types of attention mechanism: Multi-Head Attention (MHA), Multi-Query Attention (MQA), Group-Query Attention (GQA) and Multi-Head Latent Attention(MLA). Multi-Head Attention is a mechanism that allows the model to focus on different parts of the input data simultaneously by employing multiple attention heads, which is the most widely used self-attention mechanism in the Transformer models. Multi-Query Attention simplifies multi-head attention by using a single shared query across all heads but allowing different key and value projections. This reduces the complexity in both space and time. Group-Query Attention is a variant of multi-head attention that reduces computational complexity by sharing query representations across multiple heads, while allowing separate key and value representations. The idea is to use fewer query groups but still preserve a level of diversity in the attention mechanism. Multi-Head Latent Attention achieves better results than MHA through low-rank key-value joint compression, and requires much less Key-Value(KV) Cache.

Figure 2(b)① shows the changing situation of choosing three self-attention mechanisms during these time periods from 2022 to 2024. We can see that MHA is gradually being phased out and replaced by GQA.

2) *The type of feed-forward neural network.* Feed-forward network can be summarized into two types: the Standard FFN and the Gated FFN. The Standard FFN is a two-layer neural network with a activation function. The Gated FFN adds an additional gate layer.

The Figure 2(b)② shows the changing situation of type of FFN during these time periods from 2022 to 2024. It shows that Standard FFN is gradually being phased out and replaced by Gated FFN.

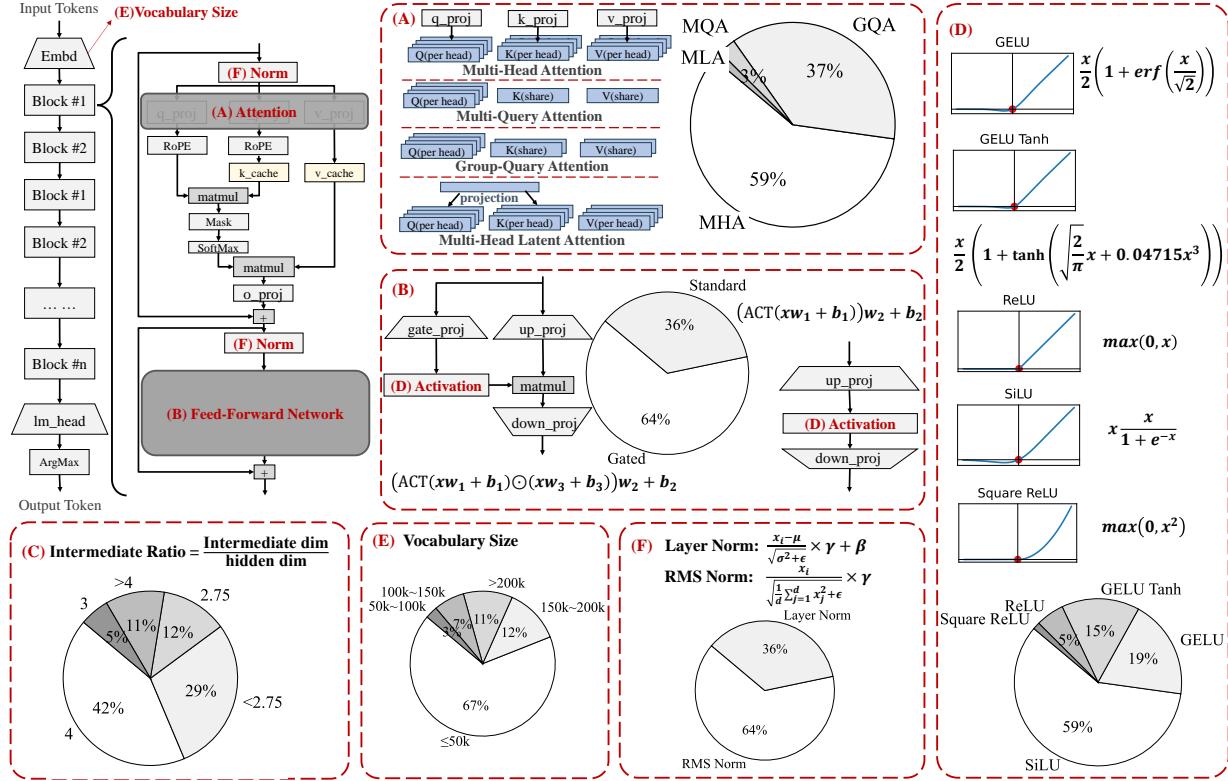
3) *The intermediate ratio of the feed-forward neural network.* The intermediate ratio of FFN is the ratio of the intermediate dimension to the hidden dimension. Figure 2(b)③ shows that the intermediate ratio of the Standard FFN is commonly set to be 4, while the intermediate ratio of the Gated FFN is rather diversified ranging from 2 to 8.

4) *The activation function of the feed-forward neural network.* There are 4 main kinds of activation functions used in FFN: ReLU (Rectified Linear Unit), GELU (Gaussian Error Linear Unit), $GELU_{\tanh}$, SiLU (Sigmoid Linear Unit). Observed from Figure 2(b)④ , the activation function of FFN was mostly ReLU in 2022, and then changed to GELU and its variants in 2023. For those released in 2024, SiLU becomes the dominant type.

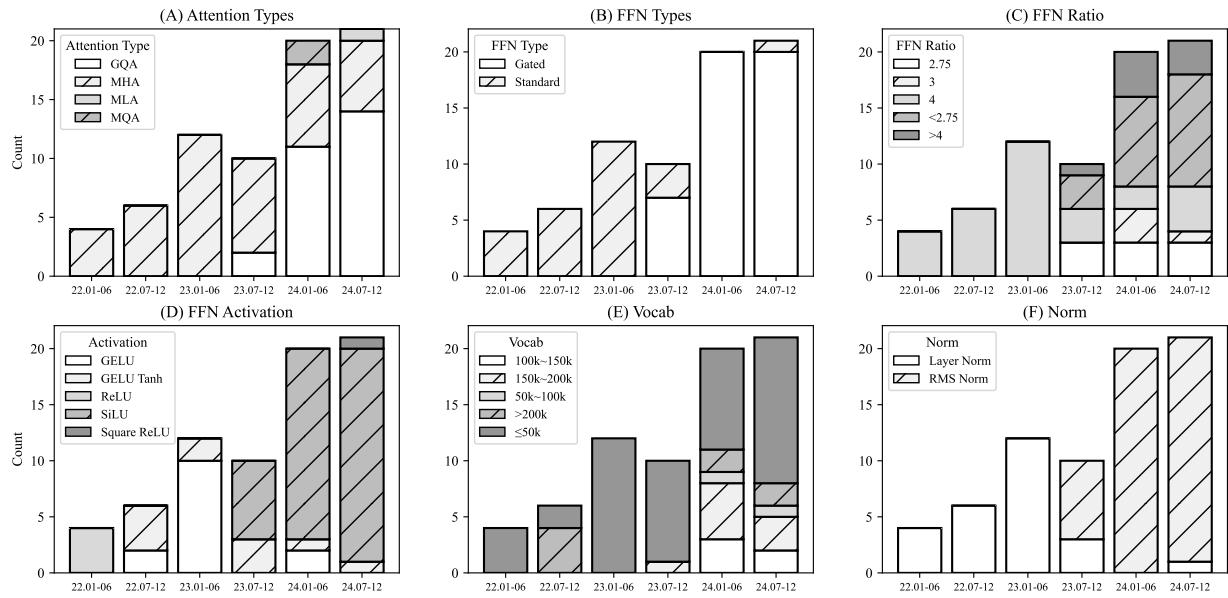
5) *The type of layer normalization.* There are two main types of layer normalization: LayerNorm and RMSNorm. The Figure 2(b)⑤ shows the changing situation of type of the type of layer normalization during these time periods from 2022 to 2024. layer normalization is gradually being phased out and replaced by RMS normalization.

6) *The vocabulary size.* The vocabulary size is the total number of unique tokens that an SLM can recognize. The Figure 2(b)⑥ shows the changing situation of the vocabulary size during these time periods from 2022 to 2024. We can see that the vocabulary size of the model is gradually increasing. The vocabulary of the latest models is often larger than 50k

Model architecture innovations. While the vanilla transformer architecture has been well recognized for its scaling ability, there still exist a few architecture-level innovations in the tested SLMs, namely parameter sharing and layer-wise parameter scaling.



(a) The architecture.



(b) Architecture distribution.

Figure 2: The architecture analysis of the SLM, highlighting 6 configurations: attention type, FFN type, FFN ratio, FFN activation, vocabulary size, and normalization type. (a) presents the overall structure of the SLM, and the categorizations with usage frequency of the 6 configurations; (b) analyzes the concrete selections of six configurations over time.

1) Parameter Sharing. Parameter Sharing is a technique used in large language models to reuse the same set of weights across different layers or components of the network. This approach allows the model to significantly reduce the number of parameters, leading to more efficient training and inference, while maintaining performance.

Embedding-lm_head sharing. Sharing the weights of the embedding with the final lm_head layer is the most common weight sharing technique. It is the sharing of the word embedding layer and has nothing to do with the rotary position encoding. Models such as Gemma, and Qwen all used this sharing technique.

layer-wise attention/FFN sharing. In this approach, the same set of weights is reused across multiple layers of the model. This is commonly seen in SLM/LLM, where all the transformer layers share the same parameters. For example, MobiLLaMA shares the weights of the FFN of all the transformer blocks; MobileLLM shares the weights of the Attention and FFN of two adjacent transformer blocks.

2) Layer-wise parameter scaling. This technique was proposed and used by OpenELM. Traditional SLMs use the same configuration for each transformer layer in the model, resulting in a uniform allocation of parameters across layers. Unlike these models, each transformer layer in OpenELM has a different configuration (e.g., number of heads and feed forward network dimension), resulting in variable number of parameters in each layer of the model. This lets OpenELM to better utilize the available parameter budget for achieving higher accuracies.

3) Nonlinearity compensation. PanGu- π analyzes the state-of-the-art language model architectures and observes the feature collapse problem. PanGu- π adopts two techniques for nonlinearity compensation of language model to solve the feature collapse problem. The series activation function is adapted to FFN, and the augmented shortcuts are integrated into MHA, which effectively introduces more nonlinearity into the Transformer architecture.

Insights: We make two key observations in SLM architectures.

- As of August 2024, a typical SLM architecture tends to use group-query attention, gated FFN with SiLU activation, an intermediate ratio of FFN between 2 and 8, RMS normalization, and a vocabulary size larger than 50K. However, the choice of such settings is mostly empirical, without strict and public validation on the superiority of such model’s capacity. Instead, the architecture innovations have relative larger impacts on the runtime performance on devices, as will be shown in §4.
- The innovations to the transformer architecture is limited in nowaday SLMs. For the few that did contribute architectural innovation (except embedding-lm head sharing), we do not observe strong evidence showing them being significantly superior to the vanilla transformer, and neither are them being generally adopted or studied across different research groups or companies. The significance of those innovations remain to be explored and validated.

2.3 Training Datasets

We investigate how the open-sourced pre-training datasets are used in training the SLMs. Overall, we find 12 such datasets being used:

- The Pile [30] (825B tokens): a combination of smaller corpora in various domains.
- FineWeb-Edu [69] (1.3T tokens): a collection of educational text filtered from FineWeb.
- StarCoder [49] (35B tokens): Python tokens.
- Cosmopedia [13] (25B tokens): a dataset of synthetic textbooks, blogposts, stories, posts and WikiHow articles generated by Mixtral-8x7B-Instruct-v0.1.
- RefinedWeb [70] (5T tokens): despite extensive filtering, high-quality data extracted from the web remains plentiful, obtained from CommonCrawl.
- RedPajama [23] (1.2T tokens): includes over 100B text documents coming from 84 CommonCrawl snapshots and processed using the CCNet pipeline.
- Dolma [76]: a English corpora, which is deduplicated inner corpus and across corpus using MinHash algorithms.
- WuDaoCorpora [100] (4T tokens): a super large-scale Chinese corpora, containing about 3T training data and 1.08T Chinese characters.
- RoBERTa [53] CCNewsV2: containing an updated version of the English portion of the CommonCrawl News dataset.
- PushShift (.io Reddit [12]: a social media data collection, analysis, and archiving platform that since 2015 has collected Reddit data and made it available to researchers.

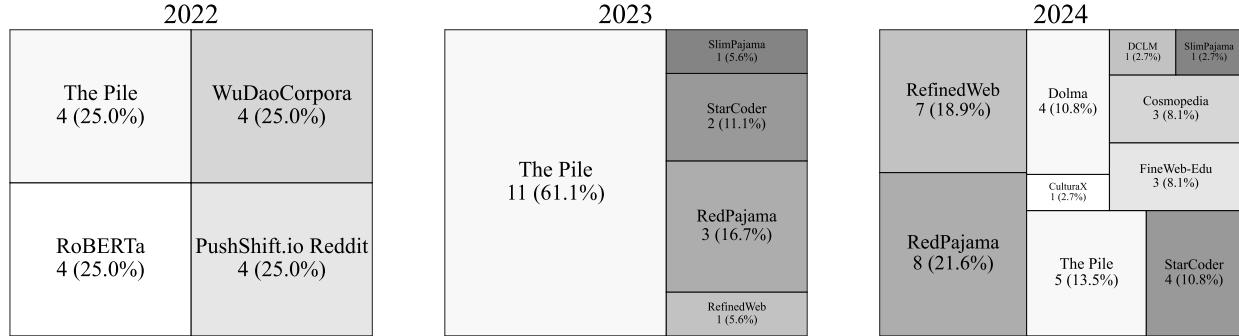
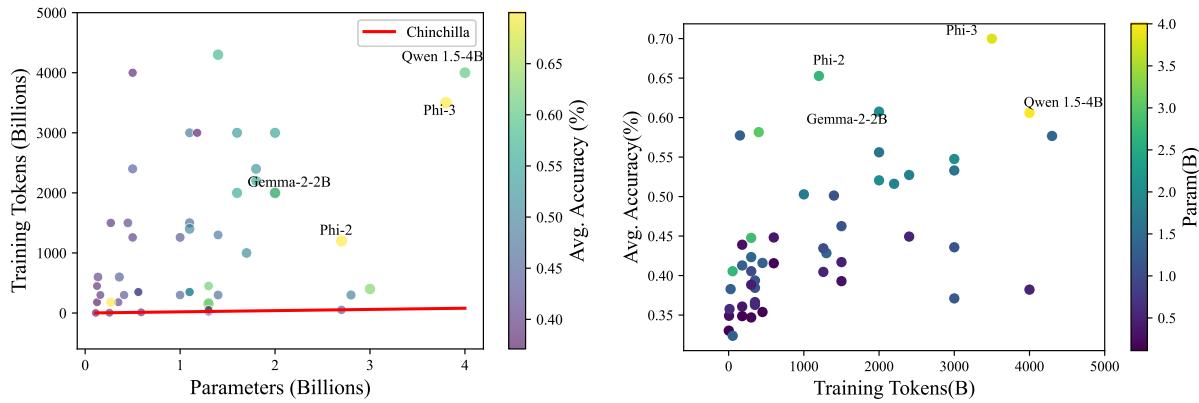


Figure 3: The usage frequency of each open-source pre-training dataset from 2022 to 2024



(a) The relationship between Training Tokens and Parameters.

(b) The influence of Training Tokens on Accuracy

Figure 4: The relationship between the number of training tokens, the number of model parameters, and the model accuracy. Here, the “accuracy” is averaged across all benchmarks in Table 2. (a) The relationship between the number of training tokens and model parameters size. According to scaling law(Chinchilla), that SLMs are often over-trained for better performance at deployment stage. (b) The influence of the number of training tokens on the model accuracy.

- DCLM-baseline [47] (1.35T tokens): a standardized corpus extracted from Common Crawl, effective pre-training recipes based on the OpenLM framework, and a broad suite of 53 downstream evaluations.
- CulturaX [65] (6.3T tokens): a substantial multilingual dataset in 167 languages.

The usage preference of pre-training datasets. We then conducted statistics on the usage frequency of the datasets for training SLM from 2022 to 2024. The results are illustrated in Figure 3. It shows that The Pile is the most widely used pre-training dataset especially in 2022 and 2023; yet more recently, more such datasets are proposed and the choice becomes diversified. In fact, The Pile has been abandoned in pre-training SLMs recently, and datasets such as “RefinedWeb” and “RedPajama” have gradually been widely used. It shows the active research and engineering efforts in constructing pre-training datasets with better quality.

Comparing the quality of pre-training datasets. We also studied the open-sourced pre-training datasets quality based on the performance of SLMs trained on them. We classified the SLMs in the past three years into groups of less than 0.5B, 1B, 2B, and 3B according to the number of parameters, and sorted them based on the average accuracy (the average of the accuracy of the two types of tasks, Commonsense reasoning/understanding and Problem solving, as will be shown in later section) of the four groups of models to explore how to select datasets to improve the average accuracy. The results are shown in Table 2. We notice that two recently released datasets, DCLM and FineWeb-Edu, show superior performance compared to others. One common feature of the two datasets is the adoption of model-based data filtering. In addition, coding data is often included in the SLM pretraining datasets such as StarCoder, even though coding ability is not a focus of SLMs deployed on devices. This is likely attributed to the common belief that coding data can help improve the model reasoning ability [106].

The number of training tokens vs. the size of model parameters. The number of parameters in SLM models and the amount of data used for training (the number of tokens) are closely related, with the Chinchilla law [37] suggesting that the optimal ratio between the number of model parameters and training tokens should be around 20

Subcaption	Model	Date	Tokens(B)	Datasets	Acc(Avg) ↓
<1B	SmolLM-360M	24.07	600	FineWeb-Edu ^b ,StarCoder,Cosmopedia ^a	0.448
	OpenELM-450M	24.04	1500	RefinedWeb, The Pile, RedPajama, Dolma	0.417
	SmolLM-135M	24.07	600	FineWeb-Edu ^b ,StarCoder,Cosmopedia ^a	0.416
	MobiLLama-0.5B	24.02	1259	RedPajama, 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-baseline ^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, RefinedWeb	0.434
	MobileLLaMA-1.4B	23.12	1300	RedPajama	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
	Bloom-1B1	22.11	350	WuDaoCorpora	0.394
	Bloomz-1B1	22.11	350	WuDaoCorpora	0.384
	Cerebras-GPT-1.3B	23.03	26	The Pile	0.383
	StableLM-2-zephyr-1.6B	24.01	2000	RefinedWeb, RedPajama, The Pile, StarCoder, CulturaX	0.556
1.5B–2B	SmolLM-1.7B	24.07	1000	FineWeb-Edu ^b ,StarCoder,Cosmopedia ^a	0.503
	StableLM-zephyr-3B	23.11	400	RefinedWeb, RedPajama, The Pile, StarCoder	0.582
	Pythia-2.8B	23.03	300	The Pile	0.448
	OPT-2.7B	22.05	180	RoBERTa, The Pile, PushShift.io Reddit	0.439
2.5B–3B	Cerebras-GPT-2.7B	23.03	53	The Pile	0.405

Table 2: Classify according to the model parameter quantity and sort in descending order according to average normalized accuracy. Acc(Avg) is the average of the accuracies of the two types of tasks, Commonsense reasoning/understanding and Problem solving. **a** indicates that this dataset is generated by LLM. **b** indicates that this dataset has been filtered by LLM.

(e.g., a 1B model with 20B tokens). We have statistically analyzed the number of training tokens used by SLMs under 4B parameters from 2022 to 2024, as shown in Figure 4(a). Generally, the larger the model, the greater the number of tokens used for training, and more recent models tend to have more training tokens. A notable observation is that SLMs are trained on much large number of tokens (typically over 1.5T) than what is suggested by the Chinchilla law, regardless of their parameter sizes. In some cases, smaller SLMs are trained on even more data than the larger SLMs (e.g., Qwen2-0.5B with 12T tokens compared to Qwen2-1.5B with 7T tokens). This indicates that these SLMs are significantly "over-trained". The rationale behind this approach is to deploy powerful SLMs on resource-constrained devices by using more training-time FLOPs. Though, SLMs are known to have performance saturation issue with over-training [31].

The amount of training tokens vs. model accuracy. Figure 4(b) shows the relationship between the number of training tokens and the accuracy of the model. In general, there is a positive correlation between the two metrics, especially for those with less than 700B tokens. However, the correlation is weak, since the data quality often outweighs the impacts of more training tokens, especially when the training tokens exceed 1T.

Insights: We make two key observations in SLM training datasets.

- Data quality is crucial to SLM capability, which receives increasing attentions in recent SLM research. The importance of data quality to the final SLM capability typically outweighs the data quantity and model architecture configurations. A notable trend of dataset research is using model-based filtering, which result in two state-of-the-art open-sourced pre-training datasets: FineWeb-Edu (1.3T/5.4T) [69] and DCLM-baseline (4T) [47]. SLMs trained on these two datasets have achieved competitive performance to those trained on closed datasets, which have significantly advanced the SLM research reproducibility.

- Recent SLMs are trained over large amount of tokens (typically $>1.5T$), disregarding their parameter sizes. In some cases, smaller SLMs are trained over even more data (e.g., Qwen2-0.5B at 12T tokens but Qwen2-1.5B at 7T tokens). It also means those SLMs are significantly “over-trained”, as compared to the Chinchilla law [37] that estimates the parameter-token ratio to be around only 20 (e.g., 1B model with 20B tokens). The incentive of such “over-training” action is to deploy powerful SLMs on resource-constrained devices through investing more training-time compute resources.

2.4 Training Algorithms

There have been a few novel training methods to improve the model capability.

Maximal Update Parameterization(μP) controls initialization, layer-wise learning rates, and activation magnitudes to ensure analytically stable training independent of a model’s layer widths. In addition to improving training stability, μP also improves the transferability of training hyperparameters from smaller to larger scale models, which permits directly using the same settings for some optimizer hyperparameters, most notably the learning rate. For example, Cerebras-GPT trains models with Maximal Update Parameterization.

Knowledge Distillation is a crucial concept in the realm of Large Language Models (LLM). It involves extracting valuable knowledge from a large and complex teacher model and transferring it to a smaller and more efficient student model. The essence of this technique is to have the student model learn to approximate the behavior and predictions of the teacher. This is achieved by minimizing the difference between their outputs. According to our statistics, LaMini-GPT and Gemma-2 adopt Knowledge Distillation.

Two Stage Pre-training Strategy is a training strategy that involves training a model in two distinct phases. During the pretraining phase, MiniCPM only uses large-scale coarse-quality pre-training data, which is abundant and can support continuous training when provided with more computational resources. During the annealing phase, we use diverse and high-quality knowledge and ability-oriented SFT data, mixed into the pre-training data. MninCPM adopts Two Stage Pre-training Strategy.

3 SLM Capabilities

3.1 Evaluation Datasets and Metrics

We used 12 datasets across three domains to evaluate the SLM performance.

- **Commonsense Reasoning Datasets:**
 - **HellaSwag** [101]: Tests narrative understanding through plausible sentence completion.
 - **TruthfulQA** [52]: Assesses the model’s ability to avoid providing false information.
 - **Winogrande** [73]: Evaluates pronoun ambiguity resolution using commonsense reasoning.
 - **CommonsenseQA** [81]: Presents multiple-choice questions requiring everyday knowledge.
 - **PIQA** [17]: Focuses on physical commonsense reasoning and object interactions.
 - **OpenBookQA** [64]: Combines scientific knowledge with commonsense for open-book science questions.
 - **BoolQ** [20]: Tests commonsense and factual reasoning with yes/no questions.
- **Problem-Solving Datasets:**
 - **ARC Easy** [21]: Contains simple science questions testing general knowledge and reasoning.
 - **ARC Challenge** [21]: Presents complex science exam questions requiring knowledge integration.
 - **MMLU** [36]: Evaluates problem-solving across diverse academic disciplines.
- **Mathematics Datasets:**
 - **GSM8K** [22]: Assesses grade-school-level mathematical reasoning skills.
 - **Minerva Math** [45]: Evaluates advanced mathematical reasoning across various topics.

We use *accuracy* as the primary evaluation metric. Accuracy measures the proportion of correct predictions to total examples. The default shown accuracy is instructed by 5 shots, as it is the most common setting in the released model. For commonsense reasoning, problem-solving, and mathematics tasks, accuracy evaluates the model’s ability to select correct options or provide accurate solutions.

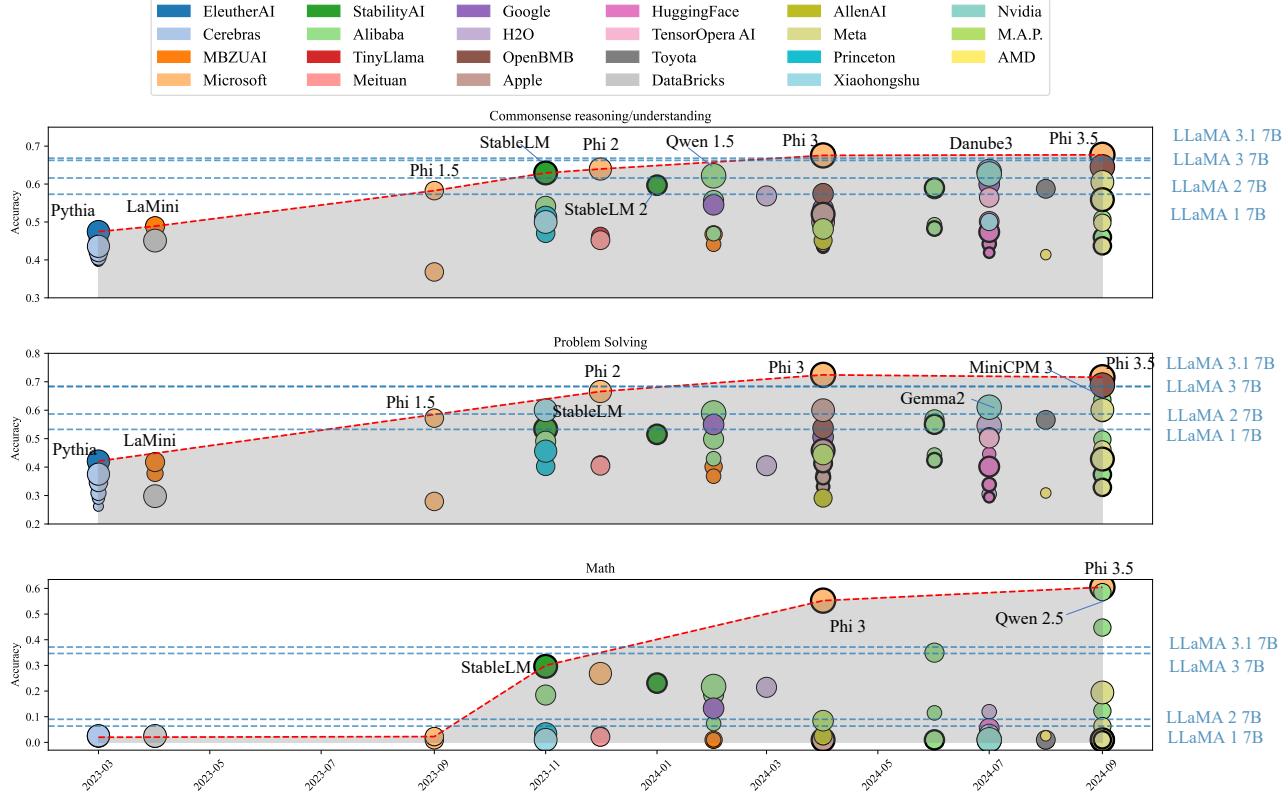


Figure 5: SLM capabilities over time. The size of the circle is proportional to the model size. Red dashed lines show the state-of-the-art model at different time, indicating the trend that SLMs are getting better over time. LLaMA-7B series models are shown in horizontal blue dashed lines for comparison. Note that Phi and StableLM series are instructed models, while others are base models.

3.2 Overall Capabilities

As shown in Figure 5, we conducted experiments on selected SLMs across three tasks—commonsense reasoning, problem-solving, and mathematics—to analyze their progress. The results show substantial performance improvements across all tasks between 2022 and 2024. Specifically, model performance improved by 10.4%, 13.5%, and 13.5% for the three tasks, respectively. In comparison, the state-of-the-art open-source LLaMA model exhibited an average improvement of only 7.5% over the same period. Notably, the Phi family, trained on closed-source datasets, outperforms all other models, achieving 67.6% in commonsense reasoning and 72.4% in problem-solving—levels comparable to the latest LLaMA 3.1 with 7 billion parameters. For example, in the mathematics task, Phi-3-mini demonstrates a substantial 14.5% lead over LLaMA 3.1. These results suggest that SLMs are rapidly closing the gap with LLMs in general reasoning tasks, although some differences remain, particularly in mathematics. Moreover, while larger parameter counts generally correlate with better performance, there are notable exceptions, such as Qwen 2, which outperforms many SLMs with 3 billion parameters despite having only 1.5 billion.

Although pioneering SLMs are trained on closed-source datasets, the gap between open-source and closed-source trained models in commonsense tasks is narrowing. For example, SmollM and DCLM-1B perform exceptionally well in commonsense reasoning (achieving 64.2% and 63.8%, respectively), thanks to high-quality datasets such as DCLM and FineWeb-Edu. However, the gap remains significant in tasks requiring complex reasoning or logic, particularly in mathematics, likely due to the lack of high-quality logic datasets.

Insights: We draw four key insights from the development of SLMs:

- From 2022 to 2024, SLMs exhibited significant performance improvements across various language tasks, outpacing the improvements of the LLaMA-7B series (1/2/3/3.1 versions). This paints a promising picture for SLMs’ potential to solve a range of downstream tasks on devices.

- The Phi family consistently achieves state-of-the-art performance across most tasks. In particular, Phi-3-mini achieves the highest accuracy as of September 2024, rivaling LLaMA 3.1 8B. While much of its superior performance may be due to careful data engineering by the Microsoft team, part of it may also be attributed to instructive tuning and potential overfitting to specific datasets [103].
- Although larger parameter counts generally lead to better performance, exceptions such as Qwen 2 1.5B demonstrate that smaller models can still excel in specific tasks.
- SLMs trained on open-source datasets are closing the gap with their closed-source counterparts in commonsense tasks. However, the gap remains significant in tasks requiring complex reasoning or logic. This underscores the need for improved datasets focused on mathematical reasoning to address this disparity.

3.3 In-context Learning Capabilities

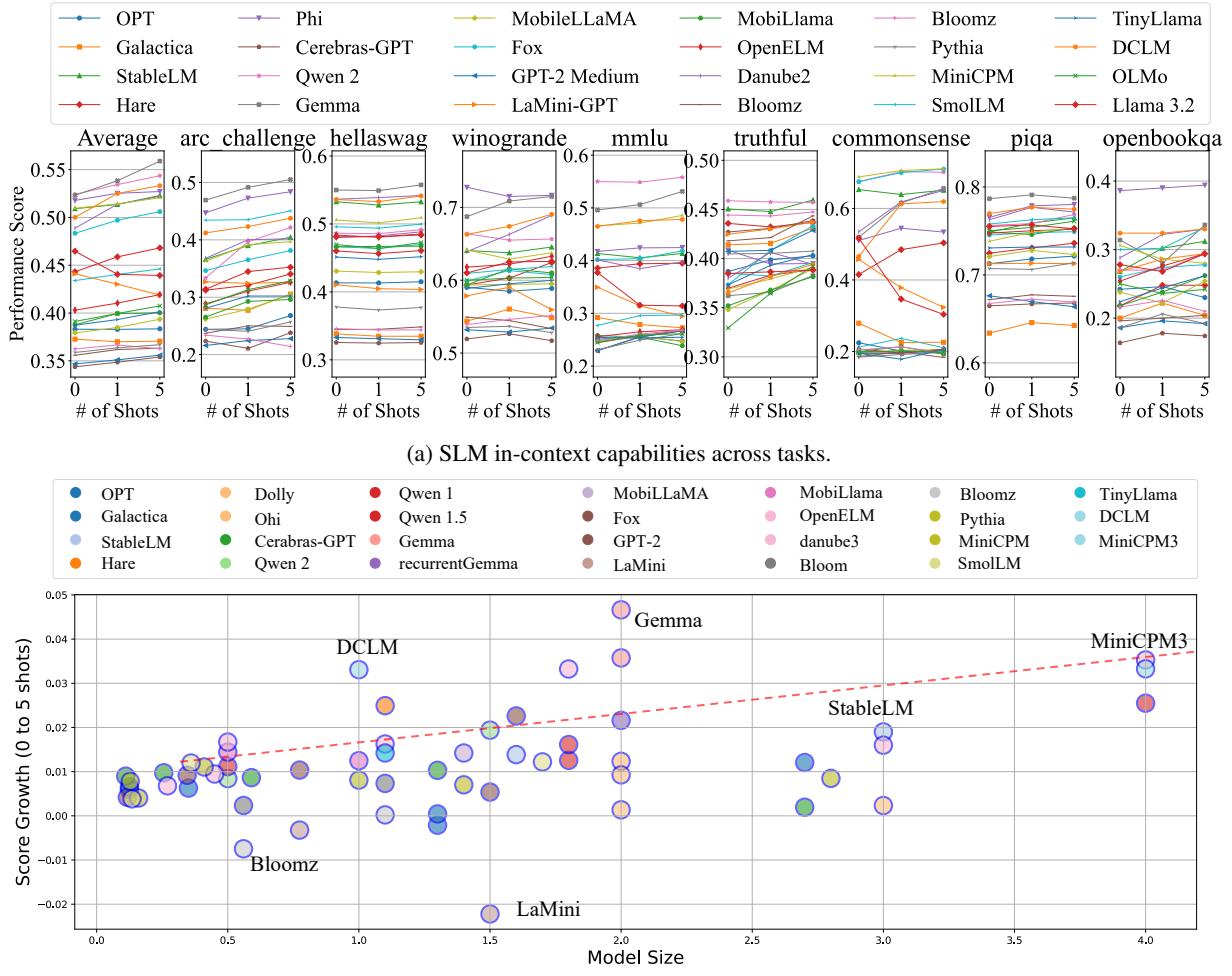


Figure 6: In-context learning performance with different tasks and models. Red line in (b) highlights the trend of the average score improvement with the increase of model size.

We conduct in-context learning experiments using various models and their 2B-parameter variants (or the closest available ones) across 8 tasks, including commonsense reasoning and problem-solving tasks. Generally, SLMs benefit significantly from in-context learning across all tasks. Exceptions include the HellaSwag and PIQA datasets, where all models perform similarly regardless of the number of in-context learning shots. These datasets are simpler and do not benefit as much from in-context learning as more complex datasets, such as ARC Challenge. On average, in-context learning with 5 shots improves the performance of zero-shot SLMs by 2.1% across all tasks. The only notable

exception is LaMini, which shows a decrease of over 2% in performance. We hypothesize that this model may be overfitting the training data, and additional context shots introduce noise. Among the models, Gemma 2 exhibits the most significant improvement, with a 4.8% increase in accuracy. Interestingly, we observe that as model size increases, the in-context learning capability of SLMs is enhanced.

Insights: We draw two key insights from the in-context learning capacity of SLMs:

- Generally, most SLMs encompass certain levels of in-context learning ability. However, such ability varies across different tasks: almost all SLMs benefit significantly from in-context learning in arc_challenge task while certain tasks show mere benefit from in-context learning across all the models, such as hellaswag and piqa.
- Larger models tend to exhibit stronger in-context learning capabilities compared to their smaller counterparts. Some small SLMs even show a decrease in performance with in-context learning.

3.4 Long Context Capabilities

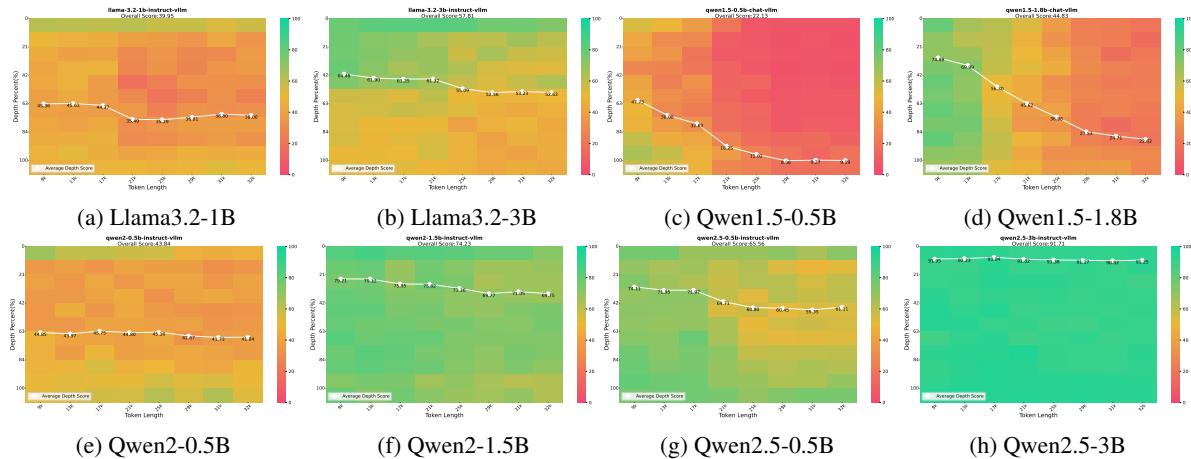


Figure 7: Needle In A Haystack

We used Needle-In-A-Haystack provided by OpenCompass to explore long-context capabilities of SLMs. The tasks included Single-Needle Retrieval, Multi-Needle Retrieval, and Multi-Needle Reasoning. The scores in Figure 7 are the average of these three tasks. Different models showed large variations in performance. Small models, such as Qwen1.5-0.5B and Qwen2-0.5B, performed less effectively. Qwen1.5-0.5B achieved an average accuracy of 22.13%. Qwen2-0.5B performed slightly better, reaching 43.84%. Qwen1.5-0.5B handled shorter contexts (9k-17k) relatively well. However, its accuracy dropped sharply with longer contexts. This was especially true for middle inserted text (Depth Percent from 20% to 70%). Larger models performed much better. Llama3.2-3B had an average accuracy of 57.81%. It worked well with shorter contexts but struggled with deeper insertions when contexts exceeded 25k tokens. Qwen2.5-3B achieved an average accuracy of 91.71%. It maintained nearly perfect accuracy across all context lengths and insertion positions. This result highlights its strong ability to handle long contexts and complex scenarios.

Insights: We draw two key insights from the long context capacity of SLMs:

- Larger parameters are crucial for long-context capabilities. Small models, such as Qwen1.5-0.5B and Qwen2-0.5B, perform adequately on short-context tasks but experience a significant drop in recognition accuracy as the context length increases. In contrast, larger models, such as Qwen2.5-3B, excel with outstanding performance, maintaining near-perfect accuracy across all context lengths and insertion positions.
- “Lost in the Middle” also occurs in small models. Compared to deep or front insertions, the accuracy of middle-position text (Depth Percent 20%-70%) is significantly lower.

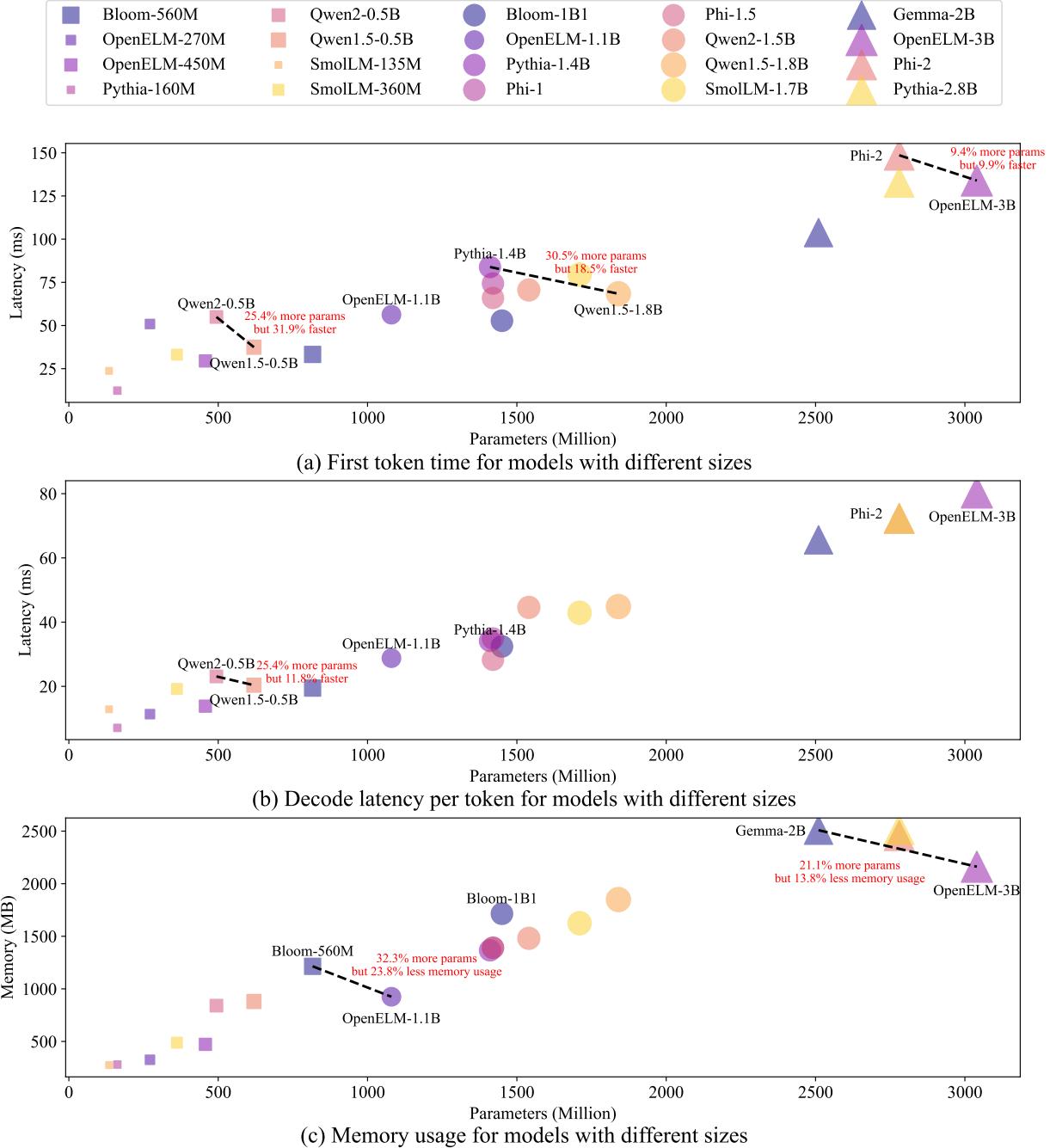


Figure 8: Latency and memory overview.

4 SLM Runtime Cost

Setup In this section, we first provide an overall analysis of the latency and memory used by models with different parameter sizes. Next, we examine the impact of quantization methods and hardware on model latency. Finally, we break down the latency and memory usage to identify the key factors affecting them in different parts of the model.

We evaluated 20 models on two types of edge devices: the Jetson Orin Module, commonly used in edge AI devices such as drones and small robots, and smartphones, which people rely on in their daily lives. The detailed specifications are shown in Table 3. All experiments on the Jetson used its GPU, while those on the smartphones were performed

Device Name	Specifications	Release Time
Jetson Orin NX 16GB	1024-core NVIDIA Ampere architecture GPU with 32 tensor cores, 16G DRAM	Feb. 2023
Pixel 7Pro	GoogleTensor G2,12G RAM	Oct. 2022
Xiaomi 12S	Snapdragon 8Gen1+,12G RAM	Jul. 2022
MEIZU 18Pro	Snapdragon 888,8G RAM	Mar. 2021

Table 3: Testing devices.

using its CPU. To eliminate the impact of inference engine implementations, we carry out all experiments using `llama.cpp`, a widely recognized open-source inference engine.

We primarily recorded metrics as model parameter, latency during the prefill and decode phases, and runtime memory usage. Due to variations in how each model officially reports its parameter counts, we relied on the parameter values obtained from `llama.cpp`. Inference is divided into two phases: prefilling and decoding. During the prefill phase, the input prompt is processed to generate a KV Cache, where multiple tokens in the prompt can be computed in parallel. For the prefill phase, we focused on first token latency, which represents the time it takes to process all tokens in the prompt. The decode phase, also known as the autoregressive phase, generates one token at a time, incorporating it into the KV Cache. Simultaneously, this token is used in predicting the next token. For the decode phase, we measured latency per token. We set a standard prompt length of 50 and a token generation length of 50 unless specified otherwise. Tests are conducted at 10-second intervals to mitigate potential thermal throttling issues. To measure larger models, we applied 4-bit quantization to all models before conducting experiments in all sections except § 4.2.1. Therefore, the latency and memory usage reported are models after 4-bit quantization. Further optimization in recent literature is thoroughly discussed in § 4.4.

4.1 Overview

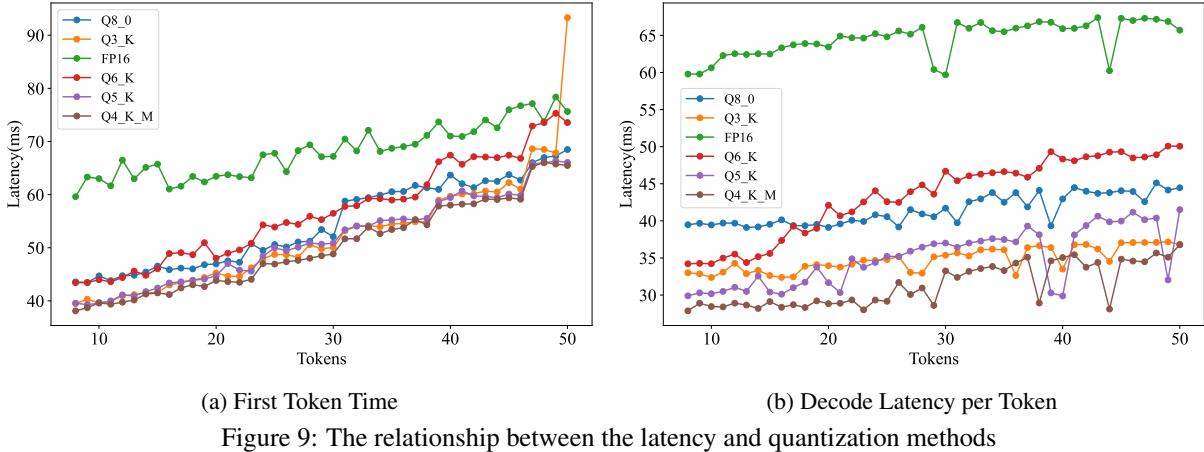
4.1.1 Inference Latency

In Figure 8, the inference latency including first token time and decode latency per token for the models ranging in size from 0.1B to 3B were measured, revealing that they can be categorized into three intervals: 0.1-1B, 1-2B, and 2-3B. The inference latency within each interval is relatively similar and aligns with the latency increase as the model size grows. For models of similar size from different architectures, the first token time during the prefill stage vary significantly. The first token time of Qwen2-0.5B is 1.46× of Qwen1.5-0.5B and is close to that of OpenELM-1.1B which has 2.18× model size. Qwen2 adopts an architecture that shares the embedding layer and the LM head, allowing more parameters to be allocated to the attention block, specifically to the attention mechanism and FFN, which are more computationally intensive and time-consuming. The latency of Pythia-1.4B is higher than that of SmollM-1.7B, Qwen2-1.5B, Qwen1.5-1.8B, and Qwen-1.8B, despite these models being larger than Pythia-1.4B. A similar phenomenon is observed in larger models: among models in the 2-3B range, Phi-2 has 1.11× latency than the larger one OpenELM-3B. To be noted, prefill stage is often regarded as the dominate phase in end-to-end LLM inference on devices, since on-device LLM tasks often involve long-context understanding for context-awareness or personalization need [91].

However, the model’s latency during the decode stage more closely follows a linear relationship with model size. The latency of Qwen2-0.5B and Qwen1.5-0.5B get close. Unlike the prefill phase, Pythia-1.4B has a lower decode latency compared to larger models. Among the 2-3B models, Gemma-2B, Phi-2, and OpenELM-3B show a trend of latency positively correlating with model size.

4.1.2 Memory Footprint

The evaluation of memory footprint in Figure 8 used `llama.cpp` on Jetson. The size of models range from 0.1B to 3B parameters and the memory footprint range from 275MB to 2456MB. Due to `llama.cpp` defaulting to allocate KV cache and compute buffer according to the maximum context length of the model, models that support longer contexts end up consuming significantly more memory than others. In our experiments, we set the maximum context length for all models to 2048 to standardize memory usage. Under the same context length, memory usage is linearly related to model size. However, some models exhibit memory usage that does not align with their size, such as Gemma-2B, Bloom-560M, and Bloom-1B1. These models have larger vocabularies compared to others: Gemma-2B has a vocabulary size of 256,000, while the Bloom series has a vocabulary size of 250,880. The OpenELM series has lower memory usage compared to models of similar parameter size for two reasons. First, it uses a vocabulary size of 32,000, smaller than the 50,000 used by most models. Second, it employs GQA, which reduces the KV cache, instead of MHA. We will explain in § 4.3.2 why vocabulary size has a significant impact on model memory usage.



Insights: We have three key insights from the overview of SLM runtime cost.

- Apart from the model size, the model architecture also impacts latency. Factors such as the number of layers, the width of the FFN, the size of the vocabulary, and whether parameters are shared play significant roles. For example, Qwen1.5-0.5B has 25.4% more parameters than Qwen2-0.5B, but runs 31.9% faster on Jetson Orin NX 16GB. The correlation is likely hardware-dependent. This indicates that SLM development shall be aligned with the hardware where it will be deployed.
 - The impacts of model architecture on inference speed is more significant at prefill stage than decode stage. This is because that the computational density in the prefill stage is higher, making it more likely to be compute-bound, while the decode stage is primarily memory-bound. Differences in model architecture can more easily affect the compute-bound scenarios; for example, wider and shallower models have higher computational parallelism.
 - Runtime memory usage is generally linearly correlated with the model’s parameter count. A few models have larger memory usage compared to others with similar parameter counts, typically due to their larger vocabulary sizes. For instance, the Bloom series has a vocabulary size of 250,880, which is 5 \times to 8 \times larger than that of most models.

4.2 Impact of Quantization and Hardware

4.2.1 Impact of Quantization

The benefits of quantization for reducing inference latency on server-side GPUs likely stem from three factors: higher computational throughput of Tensor Cores for int8 operations, reduced memory access overhead, and the decrease in heat generated by reduced memory access. On mobile devices, such as Jetson, support for int8 computation is lacking, but memory access overhead can still be effectively reduced. This reduction comes from data compression due to the lower precision of activation values and parameters, which in turn improves cache utilization.

We utilized five quantization methods to test the latency of Phi-1.5, as shown in Figure 9. Qn_K (and Qn_K_M) refer to the quantization of a model to n bits using the k -quants method with a medium (M) number of parameters, while Qn_0 specifically refers to symmetric quantization of a model to n bits. For the prefill phase, when the prompt length is relatively short, quantization can reduce latency by at least 25%. However, this benefit diminishes as the prompt length increases. When the prompt length approaches 50, the $Q6_K$ and $Q3_K$ quantization methods result in latency that is nearly identical to, or even exceeds, that of the unquantized FP16 model. On the other hand, the $Q8_0$, $Q4_K_M$, and $Q5_K$ methods provide stable performance improvements. Among these, $Q4_K_M$ performs the best, reducing latency by an average of 50%. Quantization during the decode stage delivers more consistent performance gains, reducing decode latency by up to 75% and no less than 17%. As in the prefill stage, the $Q4_K_M$ method proves to be the most effective, while $Q6_K$ remains the least efficient.

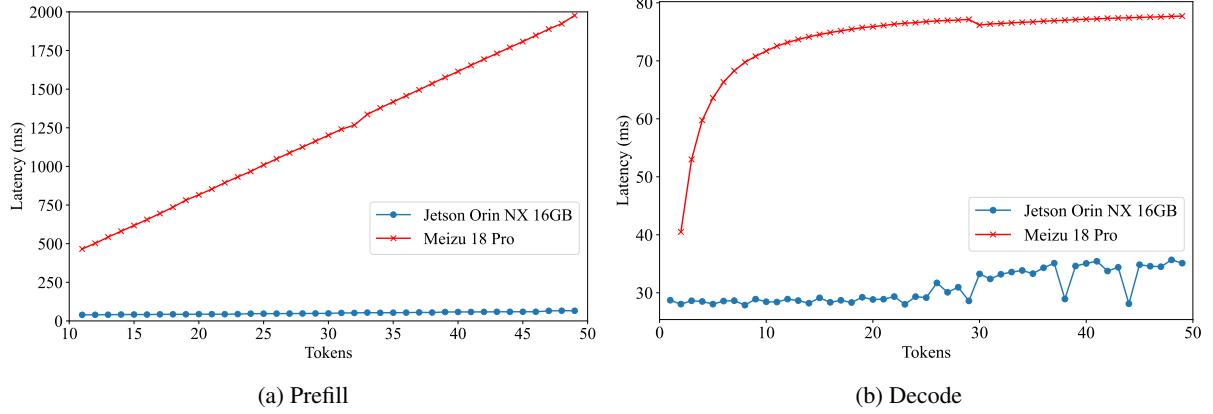


Figure 10: Latency on GPU and CPU.

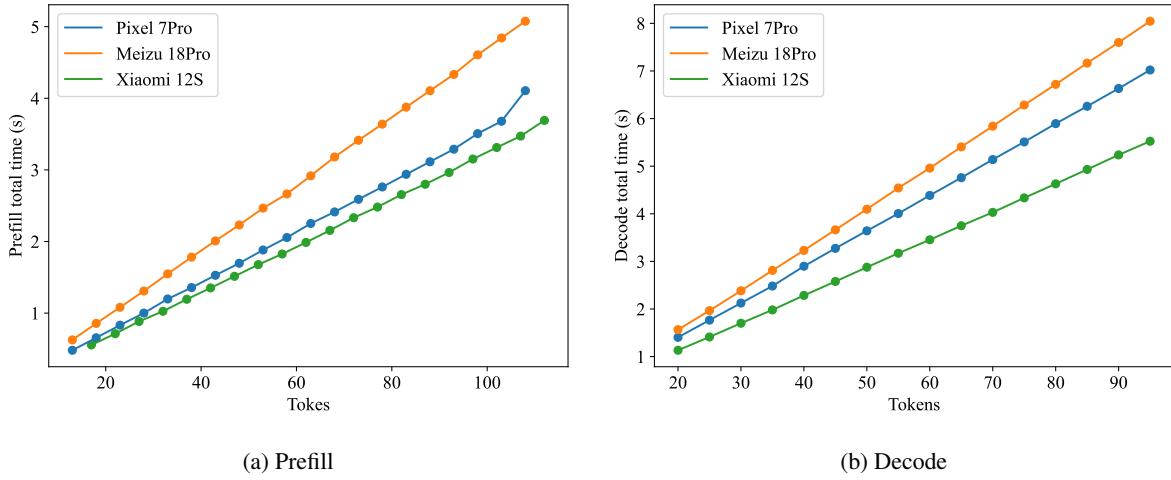


Figure 11: Latency on different smartphones.

Insights: We have two key insights about the impact of quantization methods for latency with different prompt length and output token length.

- The benefits of quantization during the decode stage are greater than those in the prefill stage. On mobile devices, quantization mainly reduces memory access overhead. Since the decode stage is more bandwidth-bound, it gains more from quantization compared to the compute-bound prefill stage.
- The benefits of quantization during the prefill stage decrease with prompt length increasing. Quantization compresses weights and kv cache. All tokens share the weight file so the benefit for each token decreases when there are more tokens in prompt. However, decode stage has more memory operation in kv cache so it still is benefited with generation tokens increasing.
- More regular quantization precision leads to better performance. Although 3-bit quantization offers a higher model compression rate, 4-bit quantization performs better in both the prefill and decode stages. The inferior performance of 3-bit quantization is due to its irregular bit-width, which lacks hardware optimization support and incurs additional overhead from data alignment and padding. As a result, despite its lower compression rate, 4-bit quantization is more efficient. Similarly, irregular 5-bit and 6-bit quantization result in inference latency that is comparable to, or even higher than 8-bit quantization, despite offering higher compression rates.

4.2.2 Impact of Hardware

We conducted tests using Bloom-1B1 on two types of edge devices: the Jetson Orin NX 16GB, which utilizes its GPU, and the Meizu 18 Pro, which relies on its CPU. During the prefill phase, for a single token, the Jetson is approximately

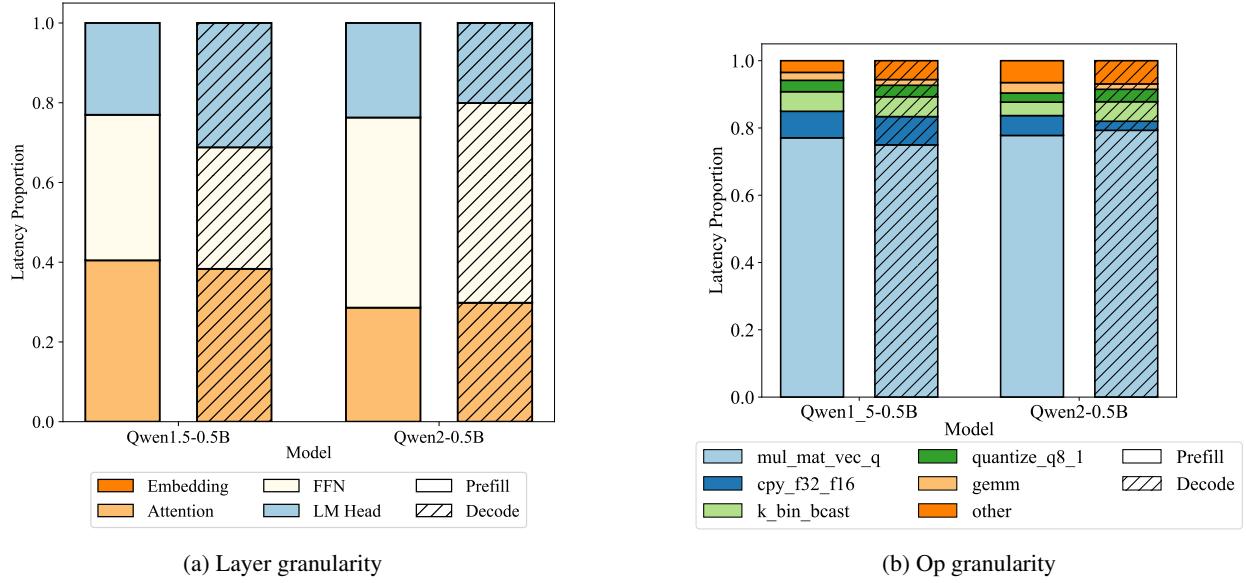


Figure 12: On-device inference latency Breakdown.

10 to 20 times faster than the Meizu 18 Pro. Both the Jetson and the Meizu 18 Pro show a linear increase in first token time as the prompt length increases, with the Jetson’s advantage becoming more obvious as the prompt length grows. During the decode phase, the latency per token increases as the number of generated tokens grows. On the Meizu 18 Pro, the latency rises sharply from 1 to 10 tokens and then levels off after 10 tokens. This initial steep rise in latency from 1 to 10 tokens is due to the temperature increase, which triggers the Dynamic voltage and frequency scaling (DVFS) or thermal throttling to adjust power consumption and frequency, thereby reducing computational efficiency. In contrast, the Jetson, benefiting from a more effective cooling system, shows noticeable fluctuations and increases in latency only after 30 tokens.

We also tested the prefill and decode times of Qwen1.5-1.8B on three smartphones. To minimize the impact of power consumption, a 60-second interval was set between each test. As shown in the figures, the latency for prefill and decode on the three smartphones increases linearly with the number of tokens. The Xiaomi 12S performed the best with the lowest latency, highlighting the efficiency of the Snapdragon 8Gen1+ chip. The Pixel 7 Pro followed, delivering competitive performance, while the MEIZU 18 Pro had the highest latency due to its older Snapdragon 888 chip and lower memory configuration.

Insights: We have two key insights about the impact of hardware for latency with different prompt length and output token length.

- GPU shows an even greater advantage over the CPU during the prefill phase. The prefill phase involves parallel processing of tokens within the prompt, whereas the decode phase generates each token sequentially. Therefore, the prefill phase has a higher degree of parallelism, making it more suitable for GPUs, which have more parallel computing units.
- The Jetson demonstrates better performance stability compared to the smartphone. Due to its relatively simple hardware structure, which facilitates better heat dissipation, the Jetson maintains more stable latency during lengthy inference tasks.
- The development of System on a Chip (SoC) generations effectively improves inference efficiency.

4.3 Latency and Memory Breakdown

4.3.1 Latency Breakdown

In Figure 12, we conducted a breakdown analysis of Qwen2-0.5B and Qwen1.5-0.5B, models with the similar size but different latency, and measured the time distribution across the Embedding, Attention, FFN(Feed-Forward Network), and LM_Head. For Qwen1.5 and Qwen2, the prefill phase is predominantly characterized by the high involvement of the Attention and FFN layers. In Qwen1.5, the Attention layer has a slightly higher proportion than the FFN layer,

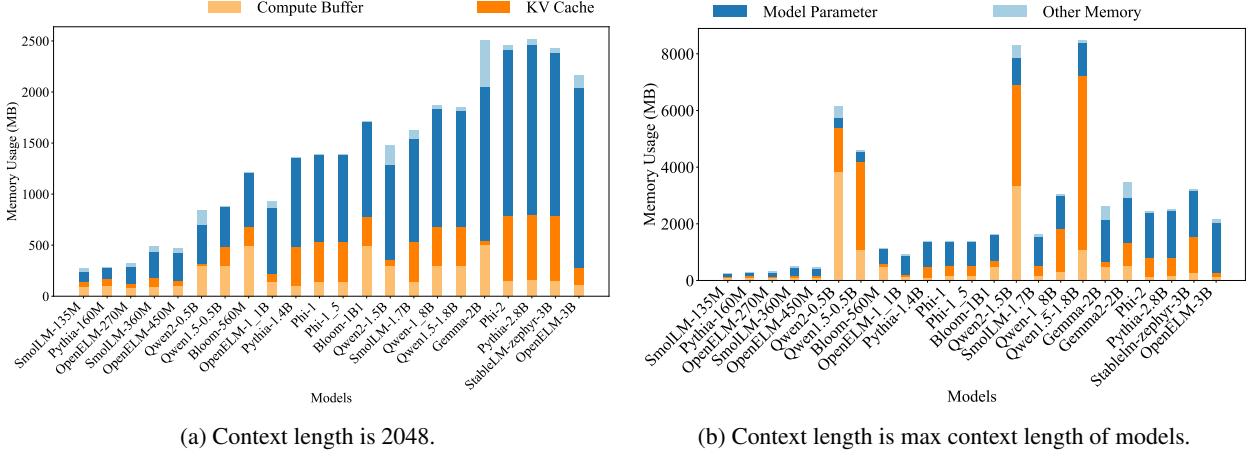


Figure 13: Memory Breakdown.

whereas in Qwen2, the FFN layer’s contribution is noticeably greater than that of the Attention layer. This is due to Qwen2 having a wider FFN layer compared to Qwen1.5. During the decode phase, the proportion of the Attention layer in Qwen1.5 increases, which could be attributed to the increased length of the KV (Key-Value) Cache. As for Qwen2, it still has longest time for FFN.

We also conducted an operator breakdown analysis on Qwen1.5-0.5B and Qwen2-0.5B. Regardless of the model or whether it is during the prefill or decode phase, the operator `mul_mat_vec_q`, which represents matrix-vector multiplication, accounts for over 80% of the time. The `mul_mat_vec_q` operator in Qwen2-0.5B has a higher proportion compared to Qwen1.5-0.5B, which may also be due to its wider FFN layer.

4.3.2 Memory Breakdown

In § 4.1.2, we found that in addition to model size, vocabulary size also has a significant impact on memory usage. In the figure, we provide a breakdown analysis of the model’s memory usage. The runtime memory is primarily consumed by model parameters, the KV cache, and intermediate variables during computation. Models with larger vocabularies require a larger compute buffer due to the need for a matrix of `hidden_size * vocabulary_size` in the Output Layer. Bloom series have a 250880 vocabulary. As Figure 13a, Bloom-560M’s compute buffer size is 492MB, which is $3.5 \times$ larger than that of the larger OpenELM-1.1B with vocabulary size of 32000. Similarly, Bloom-1B1’s compute buffer size is 496MB, which is $1.6 \times$ larger than that of the larger Qwen2-1.5B with vocabulary size of 151936. Models using GQA tend to have smaller KV Caches compared to those using Multi-Head Attention (MHA). For instance, OpenELM-3B’s KV Cache size is 164MB, which is $3.9 \times$ smaller than that of StableLM-zephyr-3B.

When the input context length is long, the sizes of the Compute Buffer and KV Cache become the primary determinants of the model’s memory usage as we seen in Figure 13b. For the Qwen2 series of models, when the context length reaches its upper limit 131072, the Compute Buffer and KV Cache occupy between 83% and 87% of the total memory. For Qwen1.5 with max context length 32768, the Compute Buffer and KV Cache occupy between 85% and 90% of the total memory.

Insights: We have two key insights for the breakdown of inference latency and memory footprint.

- `Mul_mat_vec` (matrix by vector multiplication) is the most time-consuming operations of SLM, which constitute more than 70% end-to-end inference time.
- Context length is crucial for model runtime memory usage. When context length gets to 32,000, the KV cache will take up over 80% memory.

4.4 Optimizations for On-device Deployment

Since SLMs are deployed on resource-constrained devices, numerous approaches have been developed to optimize their performance in terms of latency, memory, and other overheads. These methods can be categorized into two types: Online and Offline. Online optimization focuses on enhancing SLM performance during runtime, with key techniques including hardware-aware optimizations, model collaboration and quantization. In contrast, Offline op-

timization targets SLM improvements prior to deployment, including optimizing datasets, model architecture and knowledge distillation.

4.4.1 Online Optimization

Hardware-aware optimization requires fully leveraging the heterogeneous computing capabilities of edge devices. This involves optimizing task scheduling across CPUs, GPUs, and NPUs, as well as effectively utilizing multiple storage hierarchies, such as cache, DRAM, and flash. Earlier frameworks were primarily designed for CPU-GPU orchestration. EdgeNN [102] optimizes inference on IoT devices with a CPU-GPU architecture by improving the use of unified memory and enhancing task scheduling between the CPU and GPU, bringing an average of 3.97 \times , 3.12 \times , and 8.80 \times speedups to inference on the CPU of the integrated device, inference on a mobile phone CPU, and inference on an edge CPU device, respectively. Transformer-Lite [48] is a framework for efficiently deploying large language models on mobile phone GPUs by optimizing memory usage and computational efficiency. It utilizes techniques like model pruning, quantization, and layer fusion to reduce memory footprint and computational load, achieving up to 5.7 \times faster inference compared to CPU-based methods while maintaining model accuracy. PowerInfer [78] introduces an efficient framework for serving large language models (LLMs) on consumer-grade GPUs. It optimizes inference by partitioning the model into cold-activated neurons and hot-activated neurons, and assigns the heavy computation tasks to the GPU while offloading lighter tasks to the CPU. This results in up to 2.4 \times faster performance, reduced memory usage, and lower power consumption, making LLM inference feasible on consumer-grade hardware.

Recent systems like PowerInfer-2 [95] and mllm-NPU [91] are specifically designed for off-the-shelf modern smartphones, where the AI computing power resides in the NPU. Consequently, both frameworks focus on optimizing the CPU-NPU architecture. PowerInfer-2, as an extension of the original PowerInfer, adopts a neuron cluster approach, where matrix operations are decomposed into clusters of neurons. Based on the activation sparsity, it schedules the computation on either the NPU or CPU. This efficient allocation of resources results in up to a 27.8 \times speed improvement compared to existing state-of-the-art frameworks. Notably, PowerInfer-2 can serve a 47B parameter LLM at an impressive rate of 11.68 tokens per second, enabling fast and efficient LLM inference on smartphones without significant accuracy degradation. mllm-NPU optimizes at three levels—prompt, tensor, and block and achieves up to 22.4 \times faster prefill speed, 30.7 \times energy savings, and up to 32.8 \times speedup in end-to-end real-world applications, with the ability to process over 1,000 tokens per second for Qwen1.5-1.8B. RIPPLE [85] proposes the concept of Neuron Co-Activation, where neurons frequently activated together are linked to facilitate continuous read access and optimize data transfer efficiency. It achieves up to 5.93 \times improvements in I/O latency compared to the state-of-the-art.

For the memory optimization, LLMS [97] has proposed an effective solution. To reduce the context switching overhead when running LLMs under tight device memory constraints, LLMS employs fine-grained memory management by decoupling the memory management of the application and LLM. It utilizes globally optimized key-value (KV) cache compression and swapping techniques to efficiently manage memory resources. LLMS achieves up to 100 \times faster performance in context-switching compared to existing solutions on various edge devices.

The deployment of mixture-of-experts (MoE) models on edge devices also faces challenges in terms of huge memory usage. EdgeMoE [96] is the first on-device inference engine for MoE-based large language models, optimizing memory and computation by storing non-expert weights in device memory and expert weights in external storage, fetched only when activated. It further improves efficiency through expert-wise bitwidth adaptation and advanced expert management techniques. It reduces memory footprint by 1.05 \times to 1.18 \times compared to traditional models that store the entire model in memory. Additionally, EdgeMoE demonstrates 1.5 \times to 2.5 \times faster inference on Jetson TX2 (GPU) and Raspberry Pi 4B (CPU), with some tasks showing up to 10 \times speedup due to optimized memory and computation management. MoE Cache-Conditional [75] also chooses to keep non-expert weights resident in DRAM, while expert weights are dynamically scheduled from flash storage back to DRAM. It present on-device results demonstrating 2 \times speedups on mobile devices.

Speculative sampling is commonly used on the cloud to reduce the computational overhead of LLMs. This idea of model collaboration can similarly be applied effectively on SLMs. LLMCad [90] is an efficient on-device inference engine for generative NLP tasks on mobile devices. By combining a smaller memory-resident LLM for token generation and a larger LLM for token validation, LLMCad achieves up to 9.3 \times faster performance compared to existing engines on IoT devices and smartphones, respectively, without comprising accuracy.

Besides validating the outputs of smaller models with large models, models of different scales can also be applied to different tasks. ELMS [98] proposes a one-time neuron reordering technique for efficient sub-model generation and a dual-head compact language model for elastic mobile LLM services. ELMS improves accuracy by up to 16.83% on end-to-end traces and speeds up inference time by up to 5 \times with less than 1% switching overhead, while maintaining comparable memory usage and utilizing fewer than 100 GPU hours for model preparation.

4.4.2 Offline Optimization

Datasets play a crucial role in the accuracy of SLMs. Recent studies have focused on enhancing model performance by optimizing data utilization or generating specific types of data for training. TinyStories [26] is a synthetic dataset designed for training small language models. Trained on those child-friendly vocabulary, models as small as 28M parameters can produce coherent stories comparable to GPT-2-XL (1.5B parameters). Models trained on TinyStories require less than one GPU-day for training while exhibiting emergent reasoning and language capabilities. AS-ES learning [87] is a novel data-efficient training paradigm for improving the reasoning performance of small models by segmenting chain-of-thought data into Abstractive Segments and Extractive Segments. This method enables iterative generation without requiring additional data or modifications to the model architecture. On math word problems, AS-ES learning improves accuracy by up to 15.28%, while reducing BLEU scores, suggesting a trade-off that favors logical reasoning over memorization. For PET scan summarization tasks, it achieves up to 30.6% higher ROUGE-L scores, highlighting its effectiveness across diverse reasoning-intensive tasks. Self-AMPLIFY [14] improves small language models’ reasoning ability by generating rationales using post hoc explanation methods, eliminating the need for auxiliary models or human annotations

Model architecture optimization involves enhancing the model’s efficiency and performance by adjusting or redesigning its architecture. OnceNAS [107] is a neural architecture search method designed for on-device inference on edge devices with constrained resources. By leveraging a continuous latent space for architecture representation and integrating parameter count, latency, and accuracy as optimization objectives, OnceNAS achieves a $10.49\times$ size reduction and a $5.45\times$ speedup compared to baseline methods. Evaluations across various edge devices, including Raspberry Pi and FPGA, demonstrate superior efficiency and generalizability, making OnceNAS highly suitable for edge intelligence applications like autonomous driving and smart healthcare. Weight-Inherited Distillation (WID) [86] proposes a task-agnostic BERT compression technique that eliminates alignment losses by directly inheriting weights through structural re-parameterization with row and column compactors. WID achieves up to 49.2% parameter reduction while retaining 98.9% of BERT-base performance on GLUE benchmarks and significantly outperforms baselines like TinyBERT and MiniLM in both accuracy and efficiency, with training time reduced by over 50%.

Knowledge Distillation transfers knowledge from larger models to smaller ones through knowledge distillation, thereby improving the performance of smaller models, particularly in reasoning tasks. DISTILLM [41] proposes an efficient knowledge distillation framework, leveraging a novel skew Kullback-Leibler divergence loss and an adaptive off-policy approach to address training-inference mismatches and enhance computational efficiency. Compared to existing KD methods, DISTILLM achieves up to $4.3\times$ faster training speeds while maintaining or improving performance on OpenLLaMA2-3B. [54] explores transferring reasoning capabilities from large language models to smaller ones via chain-of-thought knowledge distillation. By fine-tuning smaller models like T5 on CoT outputs generated by LLMs such as PaLM 540B and GPT-3 175B, the authors achieve substantial improvements in reasoning tasks. For instance, T5 XXL’s accuracy on the GSM8K dataset increased from 8.11% to 21.99% when fine-tuned on CoT data generated by PaLM 540B, highlighting a 170% improvement. The approach also demonstrates robustness across arithmetic, commonsense, and symbolic reasoning tasks.

Quantization is another effective method for model compression and reducing operational overhead, but there are few methods specifically designed for edge devices. Activation-aware Weight Quantization (AWQ) [51] enables low-bit weight-only quantization for LLMs on edge devices. AWQ protects 1% of the most critical weights using per-channel scaling informed by activation patterns, achieving a $3\text{-}4\times$ speedup over FP16 implementations and reducing memory usage by $4\times$. AWQ is paired with TinyChat, a framework that further optimizes on-device inference, achieving up to $3.3\times$ acceleration on GPUs. [84] compares FP8 and INT8 formats for efficient on-device deep learning inference. It demonstrates that while FP8 may offer slightly improved accuracy for certain outlier-dominated tasks, INT8 achieves superior overall efficiency, reducing hardware costs by up to 50% and energy usage by 53%. For post-training quantization and quantization-aware training, INT8 consistently outperforms FP8 in networks with well-behaved distributions, proving more robust and computationally efficient across a range of tasks.

5 Conclusions and Future Directions

This paper makes a comprehensive survey and measurement to small language models (100M–5B parameters), including their capabilities, runtime cost on devices, and innovations. We then summarize the key insights to inspire future research on small language models. Specifically, we expect following directions worthy explorations.

Co-design and co-optimizations of SLM architecture and device processors. With given parameter size, the concrete SLM architecture still has huge impacts on the runtime speed, as discussed in §4. This includes both the basic transformer configuration (e.g., depth-width ratio, attention type, activation) and how efficiently they can be quantized for execution on integer-optimized processors like NPUs [91]. To push the limit of SLMs towards optimal

accuracy-speed tradeoff, we advocate for extreme co-design and optimizations of SLM architecture with specific device hardware, possibly searching for a speed-optimal architectures before pre-training on them.

Constructing high-quality synthetic dataset. Two recent pre-training datasets, DCLM and FineWeb-Edu, have exhibited superior performance and greatly closed the gap between SLMs trained on open/closed datasets. The key innovation of them is to use carefully trained model to filter out high-quality data portion from a large corpora. We believe that we are still at the very beginning of such synthetic data research, and the space to be explored remains huge. It is urgent to standardize a process of synthetic data curation (deduplication, filtering, mixing, evaluation, etc).

A deployment-aware Chinchilla law for model scaling. As discussed in §2.3, there is a notable trend to “over-train” SLMs on large amount of tokens as compared to what is instructed by Chinchilla law. This is because SLMs are to be deployed on resource-constrained devices, where the device memory and compute are the primary constraining factors, instead of the training-time FLOPs. Such strategy turns out to be effective to certain extent. However, the training data size cannot be scaled out infinitely, and it still remains an open question on how to determine the optimal data scaling method for SLMs. Intuitively, the decision relies on not only the training and inference cost, but also the lifecycle of SLM deployment and the economic benefits it is estimated to bring out with more training data. If sparsity is taken into consideration (e.g., MoE), the question is further complicated.

Continual on-device learning for personalization. Deployed on devices, SLMs are able to access on-device data to achieve better performance or personalization, without concerns on data leakage. There are two approaches in general. One is injecting personal data into prompts, using retrieval-augmented generation technique. This approach, however, can significantly increase the on-device cost (text embedding generation, longer prompt processing), and require the data to be stored on devices longer for retrospective query. The second approach simply uses the data to finetune the SLM, so the knowledge is embedded into the weights and the data can be discarded after finetuning. Though, such an approach faces critical challenges as well, especially the huge resource demand (memory and energy footprint) of on-device SLM training even with parameter-efficient finetuning techniques. One possible direction is to apply zeroth-order optimization to SLM finetuning [93], to avoid storing the activations in memory and be compatible with inference-time hardware accelerators (e.g., mobile NPU).

Device-cloud SLM-LLM collaboration. While our measurements demonstrate that SLM capability is fast evolving, the gap between it and the cloud-side LLM will exist. To achieve full-scale intelligence while not comprising privacy and availability (much), device-cloud collaboration will become an important research topic [19, 104]. Intuitively, the SLM can be used as a “filter” that solves the easy tasks confidently on devices, and cloud LLM can be treated as a safe guard for critical, difficult tasks. This approach bottlenecks at the capability of SLM and the decision module of what tasks are easy enough for SLMs. Does a better collaboration approach exist? Such research is challenged by the auto-regressive manner of casual language models.

Benchmarking SLMs fairly. First, SLMs are known to have systematic overfitting issue [103] to widely-used benchmarks such as GSM8k. Given most SLMs (especially those state-of-the-art ones) are trained on closed dataset, it becomes challenging to fairly compare their capability. Second, SLMs are designed to be deployed on devices, where the target tasks could differ from those hosted in clouds. There have been limited efforts in constructing a comprehensive capability and performance benchmark for SLMs. For example, when deployed on smartphones, SLMs are more likely to handle tasks that are sensitive to user data, e.g., auto-reply based on historical chatting text, or GUI context understandings. Such “ad-hoc” tasks are not included in the existing LLM benchmarks commonly used, thereby their importance is often underrepresented.

Sparse SLMs. During investigation SLMs, we find very little study of sparse SLMs, neither at architecture level (e.g., mixture-of-experts or MoE) or runtime level (e.g., activation sparsity). The reasons could be multifold. First, SLMs are supposed to have lower sparsity level as compared to LLMs [77], thereby the benefits in exploiting the sparsity for speedup or memory saving could be limited. Second, architecture-level pre-assumed sparsity method like MoE is often considered to sacrifice memory usage for less computing intensity [11, 42], which does not fit memory-constrained devices. One way to break down the memory wall for sparse SLMs is to leverage the external storage on devices, e.g., flash on smartphones. By placing “cold weights” on storage and retrieve them on demand, SLMs can be scaled out to larger capacity [7, 96, 95]. The challenges are to hide the I/O latency through careful pipeline, and keep compatible with heterogeneous hardware accelerators.

References

- [1] Google ai edge sdk for gemini nano. <https://developer.android.com/ai/aicore>, 2024.
- [2] Introducing apple’s on-device and server foundation models. <https://machinelearning.apple.com/research/introducing-apple-foundation-models>, 2024.

- [3] Alibaba. Qwen 1. https://huggingface.co/Qwen/Qwen-1_8B, 2023.11.
- [4] Alibaba. Qwen 1.5. <https://huggingface.co/Qwen/Qwen1.5-0.5B>, 2024.02.
- [5] Alibaba. Qwen 2. <https://huggingface.co/Qwen/Qwen2-0.5B>, 2024.02.
- [6] Alibaba. Qwen 2.5. <https://qwenlm.github.io/blog/qwen2.5/>, 2024.09.
- [7] Keivan Alizadeh, Iman Mirzadeh, Dmitry Belenko, Karen Khatamifard, Minsik Cho, Carlo C Del Mundo, Mohammad Rastegari, and Mehrdad Farajtabar. Llm in a flash: Efficient large language model inference with limited memory. *arXiv preprint arXiv:2312.11514*, 2023.
- [8] AllenAI. allenai/olmo-1b-hf. <https://huggingface.co/allenai/OLMo-1B-hf>, 2024.04.
- [9] AMD. Llama. <https://huggingface.co/amd/AMD-Llama-135m>, 2024.08.
- [10] Apple. Openelm. <https://huggingface.co/apple/OpenELM-270M>, 2024.04.
- [11] Mikel Artetxe, Shruti Bhosale, Naman Goyal, Todor Mihaylov, Myle Ott, Sam Shleifer, Xi Victoria Lin, Jingfei Du, Srinivasan Iyer, Ramakanth Pasunuru, et al. Efficient large scale language modeling with mixtures of experts. *arXiv preprint arXiv:2112.10684*, 2021.
- [12] Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. The pushshift reddit dataset. In *Proceedings of the international AAAI conference on web and social media*, volume 14, pages 830–839, 2020.
- [13] Loubna Ben Allal, Anton Lozhkov, Guilherme Penedo, Thomas Wolf, and Leandro von Werra. Cosmopedia, 2024.
- [14] Milan Bhan, Jean-Noel Vittaut, Nicolas Chesneau, and Marie-Jeanne Lesot. Self-amplify: Improving small language models with self post hoc explanations. *arXiv preprint arXiv:2402.12038*, 2024.
- [15] BigScience. bigscience/bloom-560m. <https://huggingface.co/bigscience/bloom-560m>, 2022.11.
- [16] BigScience. bigscience/bloomz-1b1. <https://huggingface.co/bigscience/bloomz-1b1>, 2022.11.
- [17] Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.
- [18] Cerebras. cerebras/cerebras-gpt-111m. <https://huggingface.co/cerebras/Cerebras-GPT-111M>, 2023.03.
- [19] Lihu Chen and Gaël Varoquaux. What is the role of small models in the llm era: A survey. *arXiv preprint arXiv:2409.06857*, 2024.
- [20] Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. Boolq: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019.
- [21] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. In *arXiv preprint arXiv:1803.05457*, 2018.
- [22] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.
- [23] Together Computer. Redpajama: an open dataset for training large language models, 2023.
- [24] DataBricks. databricks/dolly-v2-3b. <https://huggingface.co/databricks/dolly-v2-3b>, 2023.04.
- [25] Mateusz Dubiel, Yasmine Barghouti, Kristina Kudryavtseva, and Luis A Leiva. On-device query intent prediction with lightweight llms to support ubiquitous conversations. *Scientific Reports*, 14(1):12731, 2024.
- [26] Ronen Eldan and Yuanzhi Li. Tinystories: How small can language models be and still speak coherent english? *arXiv preprint arXiv:2305.07759*, 2023.

- [27] EleutherAI. Eleutherai/pythia-410m. <https://huggingface.co/EleutherAI/pythia-410m>, 2023.03.
- [28] Facebook. facebook/opt-125m. <https://huggingface.co/facebook/opt-125m>, 2022.05.
- [29] Facebook. facebook/galactica-125m. <https://huggingface.co/facebook/galactica-125m>, 2022.11.
- [30] Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.
- [31] Nathan Godey, Éric de la Clergerie, and Benoît Sagot. Why do small language models underperform? studying language model saturation via the softmax bottleneck. *arXiv preprint arXiv:2404.07647*, 2024.
- [32] Google. Gemma. <https://huggingface.co/google/Gemma>, 2024.02.
- [33] Google. recurrentgemma. <https://huggingface.co/google/recurrentGemma>, 2024.04.
- [34] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.
- [35] H2O.ai. h2o-danube3-4b-base. <https://huggingface.co/h2oai/h2o-danube3-4b-base>, 2024.
- [36] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2021.
- [37] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.
- [38] HuggingFace. Smollm. <https://huggingface.co/HuggingFaceTB/SmollM-360M>, 2024.07.
- [39] HuggingFace. Smollm2. <https://huggingface.co/HuggingFaceTB/SmollM2-360M>, 2024.11.
- [40] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.
- [41] Jongwoo Ko, Sungnyun Kim, Tianyi Chen, and Se-Young Yun. Distilllm: Towards streamlined distillation for large language models. *arXiv preprint arXiv:2402.03898*, 2024.
- [42] Jakub Krajewski, Jan Ludziejewski, Kamil Adamczewski, Maciej Pióro, Michał Krutul, Szymon Antoniak, Kamil Ciebiera, Krystian Król, Tomasz Odrzygóźdż, Piotr Sankowski, et al. Scaling laws for fine-grained mixture of experts. *arXiv preprint arXiv:2402.07871*, 2024.
- [43] Stefanos Laskaridis, Kleomenis Katevas, Lorenzo Minto, and Hamed Haddadi. Mobile and edge evaluation of large language models. In *Workshop on Efficient Systems for Foundation Models II@ ICML2024*.
- [44] Pierre Lepagnol, Thomas Gerald, Sahar Ghannay, Christophe Servan, and Sophie Rosset. Small language models are good too: An empirical study of zero-shot classification. *arXiv preprint arXiv:2404.11122*, 2024.
- [45] Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, et al. Minerva: Solving quantitative reasoning problems with language models. *arXiv preprint arXiv:2206.14858*, 2023.
- [46] Beibin Li, Yi Zhang, Sébastien Bubeck, Jeevan Pathuri, and Ishai Menache. Small language models for application interactions: A case study. *arXiv preprint arXiv:2405.20347*, 2024.
- [47] Jeffrey Li, Alex Fang, Georgios Smyrnis, Maor Ivgi, Matt Jordan, Samir Gadre, Hritik Bansal, Etash Guha, Sedrick Keh, Kushal Arora, Saurabh Garg, Rui Xin, Niklas Muennighoff, Reinhard Heckel, Jean Mercat, Mayee Chen, Suchin Gururangan, Mitchell Wortsman, Alon Albalak, Yonatan Bitton, Marianna Nezhurina, Amro Abbas, Cheng-Yu Hsieh, Dhruba Ghosh, Josh Gardner, Maciej Kilian, Hanlin Zhang, Rulin Shao, Sarah Pratt, Sunny Sanyal, Gabriel Ilharco, Giannis Daras, Kalyani Marathe, Aaron Gokaslan, Jieyu Zhang, Khyathi Chandu, Thao Nguyen, Igor Vasiljevic, Sham Kakade, Shuran Song, Sujay Sanghavi, Fartash Faghri, Sewoong Oh, Luke Zettlemoyer, Kyle Lo, Alaaeldin El-Nouby, Hadi Pouransari, Alexander Toshev, Stephanie Wang, Dirk Groeneveld, Luca Soldaini, Pang Wei Koh, Jenia Jitsev, Thomas Kollar, Alexandros G. Dimakis, Yair Carmon, Achal Dave, Ludwig Schmidt, and Vaishaal Shankar. Datacomp-lm: In search of the next generation of training sets for language models, 2024.

- [48] Luchang Li, Sheng Qian, Jie Lu, Lunxi Yuan, Rui Wang, and Qin Xie. Transformer-lite: High-efficiency deployment of large language models on mobile phone gpus. *arXiv preprint arXiv:2403.20041*, 2024.
- [49] Raymond Li, Loubna Ben Allal, Yangtian Zi, Niklas Muennighoff, Denis Kocetkov, Chenghao Mou, Marc Marone, Christopher Akiki, Jia Li, Jenny Chim, Qian Liu, Evgenii Zheltonozhskii, Terry Yue Zhuo, Thomas Wang, Olivier Dehaene, Mishig Davaadorj, Joel Lamy-Poirier, João Monteiro, Oleh Shliazhko, Nicolas Gontier, Nicholas Meade, Armel Zebaze, Ming-Ho Yee, Logesh Kumar Umapathi, Jian Zhu, Benjamin Lipkin, Muh-tasham Oblokulov, Zhiruo Wang, Rudra Murthy, Jason Stillerman, Siva Sankalp Patel, Dmitry Abulkhanov, Marco Zocca, Manan Dey, Zhihan Zhang, Nour Fahmy, Urvashi Bhattacharyya, Wenhao Yu, Swayam Singh, Sasha Luccioni, Paulo Villegas, Maxim Kunakov, Fedor Zhdanov, Manuel Romero, Tony Lee, Nadav Timor, Jennifer Ding, Claire Schlesinger, Hailey Schoelkopf, Jan Ebert, Tri Dao, Mayank Mishra, Alex Gu, Jennifer Robinson, Carolyn Jane Anderson, Brendan Dolan-Gavitt, Danish Contractor, Siva Reddy, Daniel Fried, Dzmitry Bahdanau, Yacine Jernite, Carlos Muñoz Ferrandis, Sean Hughes, Thomas Wolf, Arjun Guha, Leandro von Werra, and Harm de Vries. Starcoder: may the source be with you! 2023.
- [50] Xiang Li, Zhenyan Lu, Dongqi Cai, Xiao Ma, and Mengwei Xu. Large language models on mobile devices: Measurements, analysis, and insights. In *Proceedings of the Workshop on Edge and Mobile Foundation Models*, pages 1–6, 2024.
- [51] Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. Awq: Activation-aware weight quantization for on-device llm compression and acceleration. *Proceedings of Machine Learning and Systems*, 6:87–100, 2024.
- [52] Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*, 2022.
- [53] Yinhan Liu. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.
- [54] Lucie Charlotte Magister, Jonathan Mallinson, Jakub Adamek, Eric Malmi, and Aliaksei Severyn. Teaching small language models to reason. *arXiv preprint arXiv:2212.08410*, 2022.
- [55] M.A.P. Ct-llm. <https://huggingface.co/m-a-p/CT-LLM-Base>, 2024.04.
- [56] MBZUAI. Mbzuai/lamini-gpt-774m. <https://huggingface.co/MBZUAI/LaMini-GPT-774M>, 2023.04.
- [57] MBZUAI. Mobillama. <https://huggingface.co/mbzuai/MobiLlama>, 2024.02.
- [58] Meituan. Mobilellama. <https://huggingface.co/meituan/MobileLLaMA>, 2023.12.
- [59] Meta. Llama-3.2. <https://huggingface.co/meta-llama/Llama-3.2-3B>, 2024.09.
- [60] Microsoft. microsoft/phi-1. <https://huggingface.co/microsoft/phi-1>, 2023.09.
- [61] Microsoft. microsoft/phi-1.5. https://huggingface.co/microsoft/phi-1_5, 2023.09.
- [62] Microsoft. microsoft/phi-2. <https://huggingface.co/microsoft/phi-2>, 2023.12.
- [63] Microsoft. microsoft/phi-3-mini. <https://huggingface.co/microsoft/phi-3-mini>, 2024.04.
- [64] Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018.
- [65] Thuat Nguyen, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, Ryan A Rossi, and Thien Huu Nguyen. Culturax: A cleaned, enormous, and multilingual dataset for large language models in 167 languages. *arXiv preprint arXiv:2309.09400*, 2023.
- [66] Nvidia. Minitron. <https://huggingface.co/nvidia/Minitron-4B-Base>, 2024.07.
- [67] OpenBMB. Minicpm. <https://huggingface.co/openbmb/MiniCPM-1B-sft-bf16>, 2024.04.
- [68] OpenBMB. Minicpm3. <https://huggingface.co/openbmb/MiniCPM3-4B>, 2024.09.
- [69] Guilherme Penedo, Hynek Kydlíček, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, Thomas Wolf, et al. The fineweb datasets: Decanting the web for the finest text data at scale. *arXiv preprint arXiv:2406.17557*, 2024.

- [70] Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. The refinedweb dataset for falcon llm: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*, 2023.
- [71] Bo Peng, Eric Alcaide, Quentin Anthony, Alon Albalak, Samuel Arcadinho, Stella Biderman, Huanqi Cao, Xin Cheng, Michael Chung, Matteo Grella, et al. Rwkv: Reinventing rnns for the transformer era. *arXiv preprint arXiv:2305.13048*, 2023.
- [72] Princeton. Sheared-llama. <https://huggingface.co/princeton-nlp/Sheared-LLaMA-1.3B>, 2023.11.
- [73] Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.
- [74] Timo Schick and Hinrich Schütze. It’s not just size that matters: Small language models are also few-shot learners. *arXiv preprint arXiv:2009.07118*, 2020.
- [75] Andrii Skliar, Ties van Rozendaal, Romain Lepert, Todor Boinovski, Mart van Baalen, Markus Nagel, Paul Whatmough, and Babak Ehteshami Bejnordi. Mixture of cache-conditional experts for efficient mobile device inference. *arXiv preprint arXiv:2412.00099*, 2024.
- [76] Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Arthur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Harsh Jha, Sachin Kumar, Li Lucy, Xinxi Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew E. Peters, Abhilasha Ravichander, Kyle Richardson, Zejiang Shen, Emma Strubell, Nishant Subramani, Oyvind Tafjord, Pete Walsh, Luke Zettlemoyer, Noah A. Smith, Hannaneh Hajishirzi, Iz Beltagy, Dirk Groeneveld, Jesse Dodge, and Kyle Lo. Dolma: an Open Corpus of Three Trillion Tokens for Language Model Pretraining Research. *arXiv preprint*, 2024.
- [77] Jifeng Song, Kai Huang, Xiangyu Yin, Boyuan Yang, and Wei Gao. Achieving sparse activation in small language models. *arXiv preprint arXiv:2406.06562*, 2024.
- [78] Yixin Song, Zeyu Mi, Haotong Xie, and Haibo Chen. Powerinfer: Fast large language model serving with a consumer-grade gpu. In *Proceedings of the ACM SIGOPS 30th Symposium on Operating Systems Principles*, pages 590–606, 2024.
- [79] StabilityAI. stabilityai/stablelm-zephyr-3b. <https://huggingface.co/stabilityai/stablelm-zephyr-3b>, 2023.11.
- [80] StabilityAI. stabilityai/stablelm-2-zephyr*. <https://huggingface.co/stabilityai/stablelm-2-zephyr>, 2024.01.
- [81] Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019.
- [82] TensorOpera. Fox-1-1.6b. <https://huggingface.co/tensoropera/Fox-1-1.6B>, 2024.
- [83] Toyota. Dclm. <https://huggingface.co/TRI-ML/DCLM-1B>, 2024.08.
- [84] Mart van Baalen, Andrey Kuzmin, Suparna S Nair, Yuwei Ren, Eric Mahurin, Chirag Patel, Sundar Subramanian, Sanghyuk Lee, Markus Nagel, Joseph Soriaga, et al. Fp8 versus int8 for efficient deep learning inference. *arXiv preprint arXiv:2303.17951*, 2023.
- [85] Tuowei Wang, Ruwen Fan, Minxing Huang, Zixu Hao, Kun Li, Ting Cao, Youyou Lu, Yaoxue Zhang, and Ju Ren. Ripple: Accelerating llm inference on smartphones with correlation-aware neuron management. *arXiv preprint arXiv:2410.19274*, 2024.
- [86] Taiqiang Wu, Cheng Hou, Shanshan Lao, Jiayi Li, Ngai Wong, Zhe Zhao, and Yujiu Yang. Weight-inherited distillation for task-agnostic bert compression. *arXiv preprint arXiv:2305.09098*, 2023.
- [87] Nuwa Xi, Yuhang Chen, Sendong Zhao, Haochun Wang, Bing Qin, and Ting Liu. As-es learning: Towards efficient cot learning in small models. *arXiv preprint arXiv:2403.01969*, 2024.
- [88] Xiaohongshu. Minima. <https://huggingface.co/GeneZC/Minima-3B>, 2023.11.

- [89] Xiaohongshu. Minima2. <https://huggingface.co/GeneZC/MiniMA-2-1B>, 2024.07.
- [90] Daliang Xu, Wangsong Yin, Xin Jin, Ying Zhang, Shiyun Wei, Mengwei Xu, and Xuanzhe Liu. Llmcad: Fast and scalable on-device large language model inference. *arXiv preprint arXiv:2309.04255*, 2023.
- [91] Daliang Xu, Hao Zhang, Liming Yang, Ruiqi Liu, Gang Huang, Mengwei Xu, and Xuanzhe Liu. Empowering 1000 tokens/second on-device llm prefilling with mllm-npu. *arXiv preprint arXiv:2407.05858*, 2024.
- [92] Jiajun Xu, Zhiyuan Li, Wei Chen, Qun Wang, Xin Gao, Qi Cai, and Ziyuan Ling. On-device language models: A comprehensive review. *arXiv preprint arXiv:2409.00088*, 2024.
- [93] Mengwei Xu, Dongqi Cai, Yaozong Wu, Xiang Li, and Shangguang Wang. Fwdllm: Efficient fedllm using forward gradient. *arXiv preprint arXiv:2308.13894*, 2023.
- [94] Mengwei Xu, Wangsong Yin, Dongqi Cai, Rongjie Yi, Daliang Xu, Qipeng Wang, Bingyang Wu, Yihao Zhao, Chen Yang, Shihe Wang, et al. A survey of resource-efficient llm and multimodal foundation models. *arXiv preprint arXiv:2401.08092*, 2024.
- [95] Zhenliang Xue, Yixin Song, Zeyu Mi, Le Chen, Yubin Xia, and Haibo Chen. Powerinfer-2: Fast large language model inference on a smartphone. *arXiv preprint arXiv:2406.06282*, 2024.
- [96] Rongjie Yi, Liwei Guo, Shiyun Wei, Ao Zhou, Shangguang Wang, and Mengwei Xu. Edgemoe: Fast on-device inference of moe-based large language models. *arXiv preprint arXiv:2308.14352*, 2023.
- [97] Wangsong Yin, Mengwei Xu, Yuanchun Li, and Xuanzhe Liu. Llm as a system service on mobile devices. *arXiv preprint arXiv:2403.11805*, 2024.
- [98] Wangsong Yin, Rongjie Yi, Daliang Xu, Gang Huang, Mengwei Xu, and Xuanzhe Liu. Elms: Elasticized large language models on mobile devices. *arXiv preprint arXiv:2409.09071*, 2024.
- [99] Jinliang Yuan, Chen Yang, Dongqi Cai, Shihe Wang, Xin Yuan, Zeling Zhang, Xiang Li, Dingge Zhang, Hanzi Mei, Xianqing Jia, et al. Mobile foundation model as firmware. *arXiv preprint arXiv:2308.14363*, 2023.
- [100] Sha Yuan, Hanyu Zhao, Zhengxiao Du, Ming Ding, Xiao Liu, Yukuo Cen, Xu Zou, Zhilin Yang, and Jie Tang. Wudaocorpora: A super large-scale chinese corpora for pre-training language models. *AI Open*, 2:65–68, 2021.
- [101] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019.
- [102] Chenyang Zhang, Feng Zhang, Kuangyu Chen, Mingjun Chen, Bingsheng He, and Xiaoyong Du. Edgenn: Efficient neural network inference for cpu-gpu integrated edge devices. In *2023 IEEE 39th International Conference on Data Engineering (ICDE)*, pages 1193–1207. IEEE, 2023.
- [103] Hugh Zhang, Jeff Da, Dean Lee, Vaughn Robinson, Catherine Wu, Will Song, Tiffany Zhao, Pranav Raja, Dylan Slack, Qin Lyu, et al. A careful examination of large language model performance on grade school arithmetic. *arXiv preprint arXiv:2405.00332*, 2024.
- [104] Kaiyan Zhang, Jianyu Wang, Ning Ding, Biqing Qi, Ermo Hua, Xingtai Lv, and Bowen Zhou. Fast and slow generating: An empirical study on large and small language models collaborative decoding. *arXiv preprint arXiv:2406.12295*, 2024.
- [105] Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, and Wei Lu. Tinyllama: An open-source small language model, 2024.
- [106] Xinlu Zhang, Zhiyu Zoey Chen, Xi Ye, Xianjun Yang, Lichang Chen, William Yang Wang, and Linda Ruth Petzold. Unveiling the impact of coding data instruction fine-tuning on large language models reasoning. *arXiv preprint arXiv:2405.20535*, 2024.
- [107] Yusen Zhang, Yunchuan Qin, Yufeng Zhang, Xu Zhou, Songlei Jian, Yusong Tan, and Kenli Li. Oncenas: Discovering efficient on-device inference neural networks for edge devices. *Information Sciences*, 669:120567, 2024.
- [108] Zhengping Zhou, Lezhi Li, Xinxi Chen, and Andy Li. Mini-giants:” small” language models and open source win-win. *arXiv preprint arXiv:2307.08189*, 2023.