Will LLMs Scaling Hit the Wall? Breaking Barriers with Distributed Resources on Massive Edge Devices

Tao Shen [*1] **Didi Zhu** [*1] **Ziyu Zhao** [*1] **Chao Wu** [2] **Fei Wu** [1]

Abstract

The remarkable success of foundation models has been driven by scaling laws, demonstrating that model performance improves predictably with increased *training data* and *model size*. However, this scaling trajectory faces two critical challenges: the depletion of high-quality public data, and the prohibitive computational power required for larger models, which have been monopolized by tech giants. These two bottlenecks pose significant obstacles to the further development of AI. In this position paper, we argue that leveraging massive distributed edge devices can break through these barriers. We reveal the vast untapped potential of data and computational resources on massive edge devices, and review recent technical advancements in distributed/federated learning that make this new paradigm viable. Our analysis suggests that by collaborating on edge devices, everyone can participate in training large language models with small edge devices. This paradigm shift towards distributed training on edge has the potential to democratize AI development and foster a more inclusive AI community.

1. Introduction

Scaling laws (Kaplan et al., 2020; Hoffmann et al., 2022) have been fundamental to the remarkable success of foundation models, demonstrating a predictable relationship between performance and the expansion of model parameters and training data. These laws have guided the development of increasingly powerful models, from BERT (Devlin et al., 2018) to GPT-4 (OpenAI, 2023b), showing that performance improvements can be achieved through systematic scaling of both model size and training data (Brown et al., 2020; Chowdhery et al., 2022). However, the continued application of these scaling laws requires ever-increasing amounts of data and computational resources, pushing the boundaries of what is currently feasible (Patterson et al., 2021).

Public data has been the primary *fuel* driving AI development forward. This field has witnessed an exponential growth in data requirements, from the early success of MNIST (LeCun et al., 1998b) with its 70,000 handwritten digits to ImageNet's revolutionary impact with 14 million labeled images (Deng et al., 2009). This trajectory has continued with modern large language models (LLMs) like GPT (OpenAI, 2023a), LLaMA (Meta AI, 2023), and DeepSeek (Liu et al., 2024a) series, which are trained on trillions of tokens. Recent evidence that LLaMA 3.1's smallest model (8B) (Meta, 2024) trained on 15 trillion tokens, outperforms LLaMA 2's largest model (70B) (Touvron et al., 2023) trained on 2 trillion tokens (despite being $10\times$ smaller in model size, the $7\times$ increase in training data leads to superior performance), demonstrates the paramount importance of data scaling (Sun et al., 2017; Raffel et al., 2020; DeepLearning.AI, 2024). However, we are witnessing a concerning trend of data depletion, where high-quality public data sources are becoming exhausted (Lee et al., 2021; Biderman et al., 2022). Villalobos et al. argues that human-generated public text data cannot sustain scaling beyond this decade. While recent efforts advocate for training larger models with synthetic data (Chen et al., 2024a), AI-generated content may fail to yield performance improvements (Wenger, 2024), also risks polluting public data sources (Fang et al., 2024). Moreover, stricter data privacy regulations like GDPR (Parliament & of the European Union, 2016) have made data collection increasingly difficult and expensive. This looming data scarcity suggests that scaling laws may hit a wall (Hardy-White, 2024), potentially impeding further AI advancement.

Computational resources has been the primary *engine* powering AI development. Throughout AI history, major breakthroughs have been closely tied to advances in computing power, from early models requiring single CPUs (with peak performance of 1-2 GFLOPS) to modern GPU clusters. The computational demands have grown exponentially - from BERT-Large's training requiring 64 TPU v3 chips (providing 420 TFLOPS) (Devlin et al., 2018) to GPT-3's

[*]Equal contribution [1]College of Computer Science and Technology, Zhejiang University, Hangzhou, China [2]School of Public Affairs and Academy of Social Governance, Zhejiang University, Hangzhou, China. Correspondence to: Chao Wu <chao.wu@zju.edu.cn>, Fei Wu <wufei@zju.edu.cn>.

Preprint. The project page is available at `https://tao-shen.github.io/Distributed-LLM-Edges/`

training on 10,000 V100 GPUs (reaching 28,000 TFLOPS) (Brown et al., 2020), while training GPT-4 reportedly required over 25,000 NVIDIA A100 GPUs (delivering a staggering 400,000 TFLOPS) (OpenAI, 2023b). More recent models like Grok 3 push these requirements even further (xAI, 2025). However, we are approaching physical limits in single-chip performance as Moore's Law slows down (Thompson et al., 2021). While massive computing clusters can compensate for individual chip limitations, maintaining such infrastructure incurs astronomical costs - estimated at over \$100M for training GPT-4 (Sharir et al., 2020) - and poses significant environmental concerns due to their enormous energy consumption, with each training run emitting as much CO_2 as 500 cars driven for a year (Schwartz et al., 2020). Moreover, this level of computing power has become concentrated among a few tech giants, creating a monopolistic landscape that effectively excludes smaller companies and academic institutions from participating in foundational AI research (Ahmed et al., 2022). This centralization of computing resources presents a significant barrier to innovation and democratization in AI development (Thompson et al., 2022).

In this paper, we propose that leveraging massive distributed edge devices offers a promising solution to overcome both data and computing barriers in AI development. Our analysis reveals two compelling opportunities: First, edge data generated from smartphones for past 5 years are projected to reach 33.1 EB, offering fresh, diverse, and contextually rich training samples. Second, the collective computing power of edge devices - with smartphones delivering 9,278 EFLOPS for past 5 years - demonstrates the feasibility of distributed model training, as training state-of-the-art models like DeepSeek-v3 would require only about 60,723 users with edge devices working (*ideally*) in parallel to match its current training setup. Based on these insights, **we argue that leveraging these massive distributed edge devices can break barriers of data and computing wall, and everyone can participate in training large models with small edge devices.** To support this position, we first analyze the critical challenges of large language models, examining both data bottlenecks (§2.1) and computational monopolization (§2.2). We then explore the hidden potential of massive edge devices, investigating their vast untapped distributed data resources (§3.1) and computational capabilities (§3.2). Building on these insights, we investigate technical approaches for overcoming large model challenges through distributed computing architectures (§4): small language models at edges (§4.1), collaborative inference (§4.2), and collaborative training (§4.3). We then identify two critical open challenges: heterogeneous device model fusion and heterogeneous device compute sharing (§5). Finally, we discuss the societal impact like AI democratization, incentive mechanisms, and environmental benefits of this paradigm shift (§A).

2. Scaling at Risk: Challenges of Data and Computing Power

2.1. The Ceiling of Public Data

Public data for pretraining is exhausting. The rapid advancement of large language models has created an insatiable appetite for training data. Neural scaling laws establish that model performance improves predictably with data quantity—a relationship that demands exponentially growing datasets (Hoffmann et al., 2022). A canonical example is GPT-3, trained on 300 billion tokens spanning books, web content, and programming code (Brown et al., 2020). Current projections suggest dataset sizes grow at 0.38 orders of magnitude (2.4×) annually (Villalobos & Ho, 2022), implying models will require three orders of magnitude more data within a decade.

Despite the internet's vast textual resources, the total stock of high-quality human-generated text remains bounded. Recent estimates place this limit at approximately 4×10^{14} tokens (Villalobos & Ho, 2022). Villalobos et al. (2024) argues that current consumption patterns suggest exhaustion of public text data by 2028, potentially accelerated to 2026 through excessive data reuse during training (a practice termed overtraining). Therefore, the finite nature of publicly available human-generated text data is expected to become a major bottleneck for LLM scaling within the next decade. Despite the current large scale of public data, the risk of data exhaustion is rapidly approaching as data demand continues to grow (Sevilla et al., 2022).

Synthetic data has potential but faces challenges. Faced with the threat of data exhaustion, researchers have proposed various solutions, among which synthetic data generation is considered one of the most promising approaches. By leveraging LLMs to produce their own training data, researchers envision self-sustaining data ecosystems. Early successes in constrained domains like mathematics and code generation, where automated verification ensures quality, demonstrate potential (Liu et al., 2023). Recent work (Chen et al., 2024a) demonstrated that diverse synthetic data enhances the performance of LLMs during both pre-training and fine-tuning.

The adoption of synthetic data faces three fundamental challenges. First, *model collapse* occurs when models iteratively train on their own outputs, causing gradual divergence from original data distributions. This recursive process amplifies biases and reduces output diversity, ultimately degrading model performance across generations (Shumailov et al., 2023; Dohmatob et al., 2024). Second, *synthetic data quality* remains inherently unverifiable in open-domain contexts. While formal domains like mathematics allow algorithmic validation, natural language lacks objective evaluation standards. The absence of ground-truth verification creates self-referential quality assessments, compromising reliability (Alemohammad et al., 2023; Wenger, 2024). Finally,

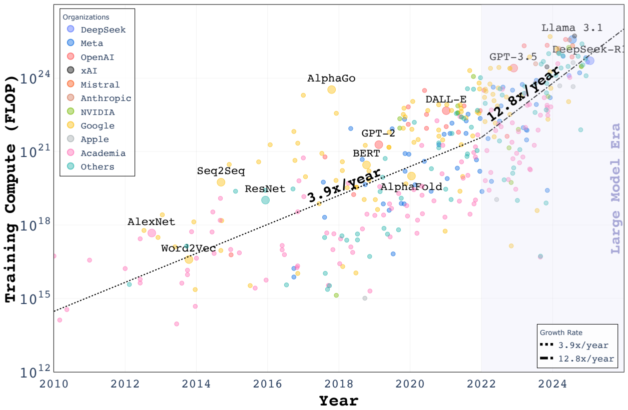

Figure 1. Trend of Computational Demand for Model Training. (Data source: Notable AI models (Hobbhahn et al., 2023)).

synthetic data struggles to replicate human *linguistic diversity*. Current methods disproportionately replicate dominant language patterns while underrepresenting cultural nuances and low-frequency expressions. This homogeneity limits their utility for training robust general-purpose models (Fan et al., 2023a).

These persistent challenges underscore that synthetic data alone cannot sustainably address the looming data scarcity crisis, compelling the research community to seek complementary strategies that transcend conventional data acquisition paradigms.

2.2. The Monopoly of Computing Resources

A few AI giants dominate the computing power. The AI computing landscape is dominated by a few major tech giants like OpenAI, Google, Microsoft, and Meta, which control powerful hardware such as GPUs and TPUs. This monopolization creates a significant barrier for smaller AI startups and research institutions, who struggle to access such advanced resources. Additionally, these companies control proprietary AI models, datasets, and software frameworks that require immense computing power, further widening the gap between the giants and smaller players. As a result, high-performance computing resources remain increasingly inaccessible to anyone outside these dominant entities.

Computational demand is growing exponentially. As large-scale AI models like GPT-4 (OpenAI, 2023b), Llama 3 (Meta, 2024), and DeepSeek-V3 (Liu et al., 2024a) surpass the trillion-parameter scale, the global AI landscape faces severe computational efficiency challenges. As shown in Figure 1, since the deep learning revolution in 2010, AI training demands have grown at a super-exponential rate of 3.9× per year—an acceleration that intensified with the adoption of the transformer architecture as the industry standard (Vaswani et al., 2017). With the advent of the era of large language models in 2022, the demand for computing power has surged even further, reaching an unprecedented growth rate of 13.4× per year. This marks a transformative shift in AI computation, where the need for computing power is expanding at an unprecedented pace, pushing the limits of existing hardware and infrastructure.

Moore's Law is slowing down. Moore's Law, which has driven the growth in computing power for decades, is slowing down as we approach the physical limits of silicon-based chip technology (Kressel, 2023). The difficulty in shrinking transistors has led to diminishing returns in computational performance. As a result, the AI industry is relying more on specialized hardware like GPUs, TPUs, and custom chips to meet growing demands. However, this shift has made high-performance hardware even more expensive and exclusive, further intensifying the gap between organizations with the resources to develop advanced AI models and those without.

Infrastructure capacity is a constraint. The rapid expansion of AI model scales and the surge in computational demand are facing dual constraints in global computing infrastructure. On one hand, bottlenecks in advanced semiconductor manufacturing severely limit the expansion rate of AI data centers. The foundry capacity for wafers at 5nm and below—such as those produced by TSMC—has already been fully booked by leading technology companies until 2026 (Staff, 2024). Moreover, the construction of new wafer fabs involves long lead times and is further constrained by the global supply chain shortages of critical equipment, such as lithography machines. On the other hand, the exponential increase in chip deployment within individual AI clusters is putting immense pressure on the already limited semiconductor manufacturing capacity, pushing the industry toward its production ceiling (ScaleFlux Research, 2024). These factors have significantly hindered the continuous expansion of computing power, making it increasingly difficult to scale AI infrastructure sustainably.

3. Scaling Beyond Limits: Opportunities from Edge Devices

3.1. Massive Data from Edges

As discussed in § 2.1, edge data represents a crucial alternative to synthetic data in addressing the challenge of data exhaustion. Edge data refers to the data generated by edge devices at or near the source of data generation, which typically remains private and localized rather than being publicly accessible. Edge devices encompass a wide range of equipment including Internet of Things (IoT) sensors, smartphones, wearables, industrial controllers, and other smart devices that process data at the network edge. The global data ecosystem exhibits significant bifurcation: publicly available data constitutes merely the tip of the iceberg, while data generated by edge devices demonstrates superior strategic value both in terms of growth rate and high quality.

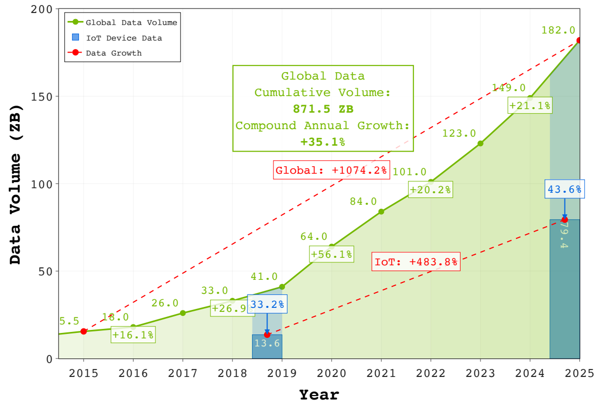

Figure 2. Global data volume from 2014 to 2025 and IoT device data volume in 2015 and 2025. (Data sources: Global data volume from (Statista global data volume, 2023); IoT device data volume from (Statista IoT device data volume, 2023).)

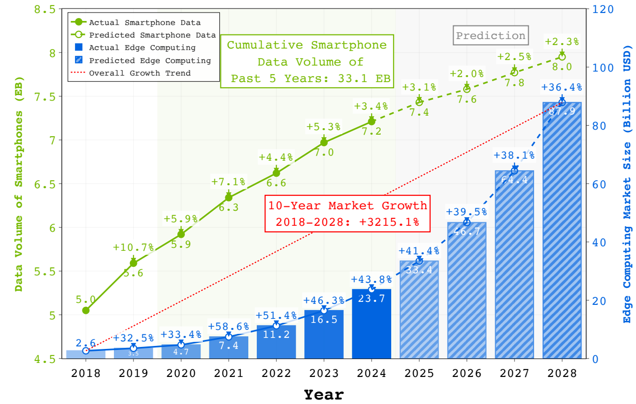

Figure 3. Smartphone data volume with edge computing market size (right) from 2018 to 2028. (Data sources: Edge computing market (Grand View Research, 2023); Smartphone data volume derived from the number of smartphones (BankMyCell, 2023).)

Edge-generated data is explosively growing. As illustrated in Figure 2, the global data volume is projected to reach 182 ZB by 2025 (Statista global data volume, 2023), with a particularly pronounced growth in edge-generated data. The data generated by IoT devices is anticipated to increase from 13.6 ZB in 2019 to 79.4 ZB in 2025 (Statista IoT device data volume, 2023), elevating its share of the global data volume from 33.2% to 43.6%. Over the period from 2015 to 2025, the global data volume exhibited a compound annual growth rate (CAGR) of 35.1%, resulting in an overall increase of 1074.2% and a cumulative total of 871.5 ZB. IoT device data experienced a growth of 483.8% from 2019 to 2025. This trend underscores the increasingly central role of edge-generated data in the global data ecosystem. Beyond IoT devices, smartphones, as a critical source of edge-side data, are also contributing to the steady rise in data volume. As depicted in Figure 3, the estimated smartphone data volume is projected to grow from 5 EB in 2018 to 8 EB by 2028 [1]. This exponential growth is closely aligned with the rapid expansion of the edge computing market, which is forecasted to surge from \$5.5 billion in 2019 to \$87.9 billion by 2028, representing a remarkable growth rate of 3215.1%. The burgeoning edge computing market has further catalyzed the generation and processing of edge-side data, reinforcing its significance in the broader data landscape.

> **Insight:** The smartphone data volume of the past 5 years (before 2025) is projected to reach approximately 33.1 EB, with unique advantages in privacy and real-time context, demonstrating the massive data potential of edge for AI model training.

Edge-generated data is of high quality. Beyond its impressive quantity, edge data possesses several qualitative advantages that make it particularly valuable for model pretraining. First, edge data exhibits superior *diversity* across multiple dimensions. It encompasses a wide variety of data types from IoT devices, mobile interactions, and personal devices, covering different domains, languages, and user behaviors. This natural diversity provides richer training signals compared to curated public datasets (Nayak et al., 2024). Second, edge data demonstrates strong *real-time capability*. Unlike public datasets, which are often updated infrequently, edge devices continuously generate fresh data with low latency (Wireless, 2023), offering more up-to-date and relevant training samples. Third, edge data offers enhanced *personalization* characteristics. It enables real-time interactions and context-aware recommendations by leveraging local data such as location and device behavior. This allows for highly relevant and adaptive personalization that can dynamically adjust to user preferences (XenonStack, 2023). Fourth, edge data exhibits enhanced *context awareness* due to the proximal positioning of edge devices to data sources, enabling the capture of fine-grained contextual information often absent in public datasets. For instance, wearable devices collect synchronized activity and location data, providing high-fidelity behavioral patterns that preserve temporal and spatial context.

In conclusion, edge data with its explosive growth [2] and superior qualitative characteristics, is a valuable resource for model pretraining. Its diversity, real-time nature, personalization, and rich context make it an ideal foundation for developing robust and adaptable large-scale models, enhancing their ability to serve real-world applications effectively.

[1] These numbers are estimated based on an average user data generation of 1 GB per device. For detailed estimation methodology, refer to Appendix C.

[2] Seagate (2019; 2020) provide a more comprehensive overview of the global data volume, but we cannot access the statistics data. We appreciate any suggestions for better statistical data sources.

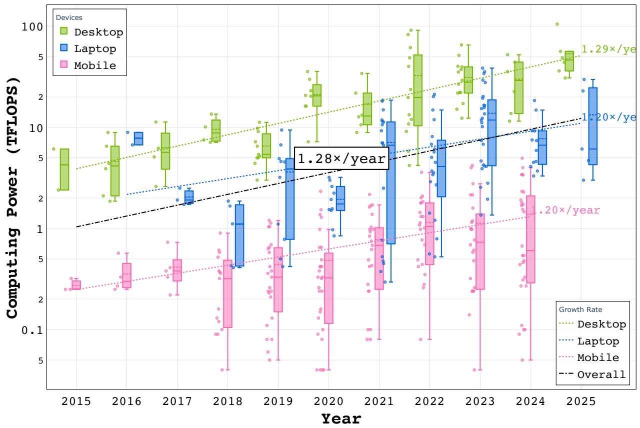

Figure 4. Edge Computing Power Evolution Trend. (Data source: (NanoReview.net, 2025)).

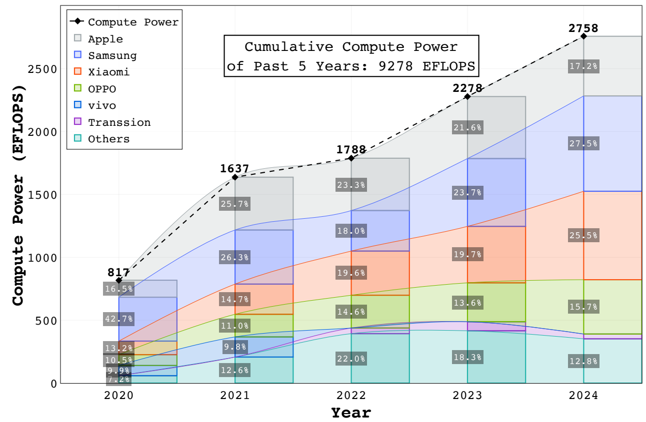

Figure 5. Smartphone Market Share and Computing Power Trends. (Data source: (Canalys, 2025)).

3.2. Massive Computing Power from Edges

Edge computing power is growing rapidly. In recent years, edge computing has experienced explosive growth in computing power, driving smart devices to evolve from single-function tools into multimodal perception and decision-making centers. For instance, as shown in Figure 4, flagship smartphones such as the iPhone 16 series, equipped with 3nm process chips, have achieved computing power exceeding 2 TFLOPS (Apple Inc., 2024), enabling local execution of complex AI tasks like real-time image enhancement and multilingual speech translation. Notably, the computing power of an individual smartphone has potentially surpassed that of laptops in the same period. The breakthroughs are even more pronounced in the desktop sector, achieving an annual computing power growth rate of 1.29×/year, surpassing that of smartphones and laptops (both 1.20×/year). These three types of devices form a differentiated growth hierarchy, collectively driving edge computing's overall computing power to expand at an annual average rate of 1.28×/year. This growth is fueled by three key technological drivers: advanced manufacturing processes (3nm technology increases transistor density by 60% (Apple Inc., 2024)), dedicated architectures (modern smartphone SoCs integrate NPUs for AI acceleration), and scenario-driven innovation (e.g., autonomous driving demands end-to-end latency of less than 100ms (NVIDIA, 2023)).

> **Insight:** The smartphone computing power of the past 5 years (before 2025) is projected to reach approximately 9278 EFLOPS, with individual flagship devices now achieving over 2 TFLOPS performance. The combined parallel computing power of approximately 30 iPhone devices (with A18 chips) can match the computational capacity of a professional AI training GPU (H100 with 59.30 TFLOPS).

Edge computing has potential for LLM training. We analyzed the performance of smartphone chips, representing typical edge devices, and estimated their overall computing power. To ensure our estimation is as accurate as possible, we based our calculations on the market share data from Canalys (2025). We then estimated the total computing power of newly produced mobile devices by averaging the chip performance of each brand. From 2020 to 2024, smartphone chip performance has seen significant improvements, with peak computing power increasing from 1.53 TFLOPS to 4.95 TFLOPS, and average computing power rising from 0.48 TFLOPS to 1.38 TFLOPS. Meanwhile, the overall computing power of mobile devices has grown from 817 EFLOPS in 2020 to 2,758 EFLOPS in 2024, and totally 9278 EFLOPS for past 5 years. This trend highlights the rapid expansion of edge computing power, which is not only essential for AI applications but also holds the potential for training complex AI models. For instance, training the DeepSeek-v3 (Liu et al., 2024a) model utilizes 2048 H100 GPUs, each providing a peak FP32 performance of 59.30 TFLOPS, resulting in a total computational capacity of 121,446.4 TFLOPS. If this workload were distributed across edge devices with a peak performance of 2 TFLOPS (e.g., mobile chips like the iPhone 16 series), approximately 60,723 users with edge devices working (*ideally*) in parallel would be required to match the computational capacity.

However, current smartphone chips are primarily optimized only for inference efficiency rather than training capabilities. *We advocate for and predict a future trajectory of edge computing where smartphone chip designs will increasingly prioritize and optimize on-device model training capabilities.* As computational power grows and distributed algorithms develop, we expect a paradigm shift enabling collaborative model training across networks of edge devices. This evolution positions the edge computing ecosystem as a critical catalyst for democratizing AI development and driving the next wave of innovations in the field.

4. Technical Advancements

4.1. Small Language Models at Edges

The first move of AI to Edge is to deploy small language models (SLMs) to edge devices (Lu et al., 2024; Wang et al., 2024a; Van Nguyen et al., 2024). This trend is driven by the growing demand for AI applications that can run directly on edge devices, motivated by needs for privacy, offline usage, and real-time processing without cloud dependence. However, edge devices have limited memory, computation, and energy resources, requiring more efficient and compact models.

SLMs leverage innovative architectures for efficient edge deployment. The classic Transformer architecture (Vaswani et al., 2017) uses self-attention mechanisms for effective sequence modeling but faces quadratic complexity challenges, with models like TinyBERT (Jiao et al., 2020) (14.5M parameters) and ALBERT (Lan et al., 2020) (12M parameters) demonstrating its effectiveness at small scales. Mamba (Gu & Dao, 2023), based on state space models, achieves linear complexity and faster inference by utilizing only the previous hidden state, as demonstrated by Zamba2-2.7B (Glorioso et al., 2024) which achieves twice the speed and 27% reduced memory overhead compared to traditional models. Hymba (Dong et al., 2024) combines both approaches by integrating attention and SSM heads within the same layer for parallel processing, with its 1.5B variant trained on DCLM-Baseline-1.0 and SmolLM-Corpus achieving 11.67 times cache size reduction while outperforming Llama-3.2-3B. The xLSTM architecture (Beck et al., 2024) modernizes LSTM with exponential gates and matrix memory cells, with models ranging from 125M to 1.3B parameters trained on 300 billion tokens from SlimPajama (Systems & Computer, 2023), consistently outperforming comparable RWKV-4 (Peng et al., 2023), Llama (Inan et al., 2023), and Mamba models across various tasks in the PALOMA benchmark (Magnusson et al., 2023). These architectural innovations demonstrate the potential for efficient and powerful language models that can run effectively on edge devices.

SLMs can be constructed through diverse methodological approaches. The construction of efficient SLMs relies on a comprehensive suite of techniques, each with specific performance trade-offs. For training SLMs from scratch, optimized MLM approaches (Devlin et al., 2019) with increased masking ratio (25% vs traditional 15%) demonstrate 2-3% performance improvements for models under 3B parameters. When deriving SLMs from existing LLMs, knowledge distillation has proven particularly effective, with response-based distillation (Sanh et al., 2019) reducing model size by 40% while maintaining 95% of the original performance. Similarly, model compression techniques like 4-bit quantization (Zhang et al., 2022) can reduce model size by 75% with only 2-3% performance drop. In architecture optimization, the Mixture of Experts approach (Fedus et al., 2022) enables 65% parameter reduction while potentially improving performance in specific tasks. Domain specialization has shown remarkable results, particularly in the medical field where 3B parameter models achieve 92% accuracy, outperforming 175B models (89%) (Fu et al., 2023). The combination of these techniques yields impressive results - a notable example is a 770M parameter model from (Shridhar et al., 2023) that combines distillation, quantization, and domain specialization to achieve 95% of the performance of a 540B model on specific tasks while requiring less than 0.15% of the computational resources. Most successful SLMs achieve their efficiency by combining multiple techniques, typically starting with knowledge distillation, applying compression methods, and finishing with domain-specific fine-tuning. Wang et al. (2024a) has provided a comprehensive survey of SLMs, and we summarize the architecture innovations and training methods in Appendix E.

Despite remarkable progress in deploying compressed models to edge devices, the current landscape remains largely confined to individual devices operating in isolation, failing to leverage the massive distributed computing power that could be achieved through collaborative training across edge devices. This represents a significant missed opportunity, as the collective computing resources of billions of edge devices worldwide remain untapped, while individual devices struggle with the computational demands of modern AI applications.

4.2. Collaborative Inference at Edges

The emergence of collaborative inference at the edge represents a significant shift in AI infrastructure, driven by the increasing need for accessible and cost-effective AI inference solutions. Traditional large-scale data centers, while powerful, often present barriers in terms of cost, energy consumption, and accessibility, providing a middle ground between edge devices and massive data centers.

Exo (Labs, 2025) is a recently popular collaborative inference framework that enables users to create AI clusters using everyday devices such as iPhones, iPads, Android devices, Macs, and Linux systems. By leveraging a peer-to-peer architecture, Exo allows these devices to work together seamlessly, effectively unifying their computational resources into a powerful AI cluster. Neurosurgeon (Kang et al., 2017) introduces a lightweight scheduler that partitions deep neural network (DNN) computations between devices and datacenters, achieving significant reductions in latency and energy consumption. MoE^2 (Jin et al., 2025) proposes a Mixture-of-Edge-Experts framework to optimize inference for large language models under energy and latency constraints. Edgent (Li et al., 2018a) enables low-latency edge intelligence through adaptive DNN partitioning and early-exit mechanisms. Galaxy (Ye et al., 2024) leverages a hybrid model parallelism approach and heterogeneity-aware planning to efficiently execute Transformer models across edge devices, reducing inference latency. These approaches collectively emphasize the synergy of device-edge collaboration for efficient deep learning inference in edge AI systems.

While collaborative inference at edge devices has made some strides, the current landscape has yet to fully capitalize on the immense potential of distributed data resources at the edge. The next frontier of research lies in transforming these devices from mere inference endpoints into active participants in the training process. This paradigm shift promises to usher in a new era of distributed AI that effectively harnesses the collective computing power and diverse data resources of billions of edge devices worldwide, fundamentally reshaping how we approach AI system development and deployment.

4.3. Collaborative Training at Edges

To harness the potential of vast amounts of data distributed across numerous devices, we envision a future where everyone can participate in training large-scale models. Federated learning emerges as a paradigm for distributed collaborative training that makes this vision possible.

Federated learning (McMahan et al., 2017b) is a practical paradigm that enables collaborative model training while preserving data privacy. Instead of collecting raw data from edge devices, which may violate privacy regulations like GDPR (Parliament & of the European Union, 2016), this approach distributes the training process across multiple devices. Each device trains on its local data and only shares model updates with the central server. This approach can effectively utilizes both computational and data resources available at edges.

Federated LLMs for fine-tuning has emerged as a critical direction in recent research of large language models, addressing the challenges of privacy preservation and resource constraints. FedLLM (Wu et al., 2024c) introduces a novel framework that enables efficient fine-tuning of large language models in a federated setting, demonstrating comparable performance to centralized fine-tuning while maintaining data privacy. FedFM (Chen et al., 2024b) tackles the critical challenges of system and statistical heterogeneity in federated learning, proposing adaptive optimization techniques that improve model convergence across diverse client devices. To address the computational constraints of edge devices, FedPET (Li et al., 2024e) employs parameter-efficient fine-tuning techniques, significantly reducing the memory and computation requirements while maintaining model performance. This approach enables even resource-constrained devices to participate in the fine-tuning process. Similarly, (Xu et al., 2024) bridges the gap between federated learning and foundation models by introducing novel techniques for efficient knowledge transfer and model adaptation in distributed settings. In domain-specific applications, FedQA (Wang et al., 2024b) demonstrates the effectiveness of federated learning for question-answering tasks, showing that models can be fine-tuned on sensitive domain-specific data while preserving privacy. These advancements are supported by open-source frameworks like Flower (Flowerlab, 2025) and FATE-LLM (Fan et al., 2023b), which provide robust platforms for implementing federated fine-tuning of large language models.

Federated LLMs for pretraining have opened up exciting new possibilities for large language model training. Rather than relying on traditional data center approaches, researchers have developed innovative geographically distributed frameworks that enable collaborative training across many devices. Notably, Prime Intellect (PrimeIntellect-ai, 2025) has launched the first decentralized training project for a 10 billion parameter model, named INTELLECT-1, which utilizes the OpenDiLoCo framework to significantly reduce communication costs between nodes. This innovative approach allows for dynamic management of computational resources across multiple locations, achieving an impressive 83% overall computational utilization while collaborating with up to 112 H100 GPUs across five countries and three continents. The model not only enhances parameter efficiency by 25 times but also demonstrates robust performance in various benchmark tests. In parallel, the Flower Lab (Flowerlab, 2025) has introduced Flower-LLM, which successfully trained a 1.3 billion parameter large language model (LLM) using novel federated learning methods. Additionally, it has also developed Photon (Sani et al., 2024), an open-source framework that provides flexible configurations for training models of different sizes, making federated LLM training more accessible to a broader range of participants and computational resources.

These frameworks underscore a shift towards decentralized AI inference and training (summarized in Appendix F), enabling researchers worldwide to contribute to advanced model development without the constraints of centralized resource control, thus paving the way for a more collaborative and inclusive AI landscape.

5. An Open Problem: How to Train Large Language Models with Small Edge Devices?

A fundamental limitation of traditional federated learning lies in its requirement for each participant to maintain and train a complete model locally. This assumption becomes particularly problematic in the context of large language models, where the computational and memory requirements far exceed the capabilities of most individual participants. For instance, in sensitive domains like healthcare (Bolton et al., 2024), multiple hospitals may wish to collaboratively

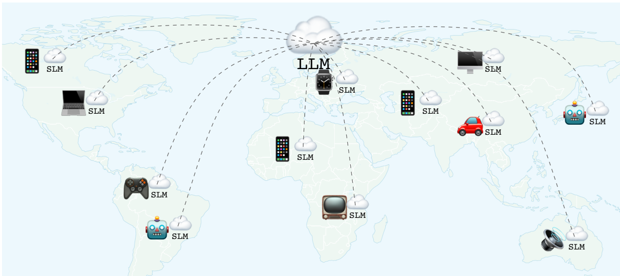

Figure 6. Train Large Language Models with Small Edge Devices

train a medical language model to leverage their collective data while maintaining data privacy. However, traditional federated learning mandates that each participating hospital possess sufficient computational resources to train the complete model locally. Despite these institutions' valuable data contributions and their motivation to enhance model capabilities through collaborative training, many hospitals lack the necessary infrastructure to participate effectively. This resource constraint significantly limits the potential for collaborative model training in critical domains where data privacy is paramount but computational resources are unevenly distributed (Kairouz et al., 2021). While some federated learning approaches allow training partial model parameters (Horvath et al., 2021; Alam et al., 2022), the enormous disparity between large language models and what edge devices can train—often orders of magnitude smaller due to inherently constrained resources—remains too vast to be effectively bridged by current FL frameworks. Therefore, we need to develop a new federated learning paradigm that enables participants to collaboratively train a large language model even under extremely limited resources (as shown in Figure 6). **It is still an open problem to train large language models with small edge devices.** Therefore, we encourage the research community to develop novel distributed collaborative computing methods in two key directions:

5.1. Heterogeneous Device Model Fusion: from Small to Large

The first direction addresses the fundamental challenge of model size disparity in federated learning. Modern large language models typically contain hundreds of billions of parameters, while edge devices have severely limited computational resources. This creates an enormous scale gap - the large target model may be hundreds or even thousands of times larger than what individual devices can handle. To bridge this gap, each edge device should run a small language model that matches its computational capacity. For example, while the central model may have 100 billion parameters, a resource-constrained mobile device might only handle a 100-million parameter model, representing a 1000x size difference. The key challenge then becomes how to effectively aggregate and fuse knowledge from these much smaller models into the large target model. We need novel techniques that can meaningfully combine insights from models operating at radically different scales while preserving the unique contributions of each small model. This requires fundamentally rethinking traditional model fusion approaches (Velasevic et al., 2023; Azizan-Ruhi et al., 2019) to handle such extreme parameter count disparities.

5.2. Heterogeneous Device Compute Sharing: from Node to Cluster

The second direction is to enable efficient compute resource sharing across heterogeneous devices by treating them as a unified compute cluster rather than independent nodes. Consider a smart home environment where multiple devices—smartphones, laptops, and desktop computers—could form a collaborative compute cluster. While each individual device has limited resources, their collective computing power could be substantial. For example, a laptop could handle intensive computational tasks, smartphones could manage coordination and lightweight processing, and desktop computers could contribute their onboard computing power. Meanwhile, other IoT devices such as smart speakers, security cameras, and vehicles could serve as data sources, providing valuable real-world inputs like voice commands, visual feeds, and environmental parameters. The language model would effectively run and train across this entire device cluster, leveraging both computing power and diverse training data from the environment. This distributed execution requires new frameworks that can intelligently decompose and distribute model computations based on each device's capabilities and current load. The system must dynamically balance workloads - when the security cameras are idle at night, they could take on additional compute tasks, while during peak usage hours, the load could shift to other devices. This requires innovations in real-time resource allocation, task scheduling across heterogeneous hardware, and efficient inter-device communication protocols to ensure the collective computing power is optimally utilized (Zhao et al., 2024).

6. Conclusion

In this position paper, we have argued that leveraging massive distributed edge devices can break barriers of data and computing wall, and everyone can participate in training large models with small edge devices. Our comprehensive analysis demonstrated the vast untapped potential of edge resources, with smartphone data volume reaching approximately 33.1 EB and a combined computing power of around 9278 EFLOPS in the past 5 years. These edge resources offer unique advantages in terms of data diversity, privacy, real-time context, and computing efficiency. This paradigm shift towards distributed training could democratize AI development and open an exciting new chapter in the scaling of foundation models.

References

Abdin, N., Chaudhuri, M., et al. Phi-3: Building better language models with more data and compute. *arXiv preprint arXiv:2404.14219*, 2024.

Ahmed, Z., Amizadeh, S., Bilenko, M., Carr, R., Chin, W.-P., Dekel, Y., Dublish, X., Greberman, M., Czerwinski, S., Hu, X., et al. Democratizing machine learning: A comprehensive survey of open-source frameworks and platforms. *arXiv preprint arXiv:2201.00118*, 2022.

AI, M. Meta releases llama 3.2. https://about.fb.com/news/2024/09/introducing-llama-3-2-1b-3b/, 2024.

AI, T. Together ai: The ai acceleration cloud. https://www.together.ai/, 2023.

AI2. Dolma: An open corpus of high-quality english text for language model pre-training. https://huggingface.co/datasets/allenai/dolma, 2023.

Alam, S., Liu, L., Yan, M., and Zhang, M. Fedrolex: Model-heterogeneous federated learning with rolling sub-model extraction. *Advances in neural information processing systems*, 35:29677–29690, 2022.

Alemohammad, S., Casco-Rodriguez, J., Luzi, L., Humayun, A. I., Babaei, H., LeJeune, D., Siahkoohi, A., and Baraniuk, R. G. Self-consuming generative models go mad. *arXiv preprint arXiv:2307.01850*, 2023.

alibaba. Federatedscope: An easy-to-use federated learning platform. https://github.com/alibaba/FederatedScope, 2024.

Allal, L. B. et al. Smollm: A small language model for everyday tasks. *arXiv preprint arXiv:2407.12290*, 2024.

Amodei, D. and Hernandez, D. Ai and compute. *OpenAI Blog*, 2, 2018.

Anderson, J., Chen, W., Liu, S., and Zhang, M. Sciencegpt: Large language models for scientific discovery and innovation. *Nature Scientific Reports*, 14:1–15, 2024a.

Anderson, M., Garcia, E., and Kumar, R. Community-driven ai development: Breaking the monopoly in large language model training. *Nature Machine Intelligence*, 6 (3):234–246, 2024b.

Apple Inc. iphone 16 pro and 16 pro max - technical specifications, 2024. URL https://www.apple.com/iphone-16-pro/specs/.

Azizan-Ruhi, N., Lahouti, F., Avestimehr, A. S., and Hassibi, B. Distributed solution of large-scale linear systems via accelerated projection-based consensus. *IEEE Transactions on Signal Processing*, 67(14):3806–3817, July 2019. ISSN 1941-0476. doi: 10.1109/TSP.2019.2917855.

Bai, J., Wang, S., Xiong, F., Hou, Z., et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

BankMyCell. How many smartphones are in the world?, 2023. URL https://www.bankmycell.com/blog/how-many-phones-are-in-the-world.

Beck, D., Ngiam, J., Mingyang, W., Le, Q. V., and Sutskever, I. xlstm: Scaling lstm to billions of parameters. *arXiv preprint arXiv:2401.02099*, 2024.

Benallal, L. et al. Smollm-corpus: An open high-quality corpus for pretraining language models. *arXiv preprint arXiv:2406.11403*, 2024.

Biderman, S., Schoelkopf, K., Weiss, A., and Noever, D. Data governance in the age of large language models. *arXiv preprint arXiv:2211.09911*, 2022.

Bolton, E., Venigalla, A., Yasunaga, M., Hall, D., Xiong, B., Lee, T., Daneshjou, R., Frankle, J., Liang, P., Carbin, M., et al. Biomedlm: A 2.7 b parameter language model trained on biomedical text. *arXiv preprint arXiv:2403.18421*, 2024.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Brown, S., Miller, D., and Wilson, J. Decentralized llm training: A path towards democratized ai development. *Proceedings of the International Conference on Machine Learning*, pp. 2145–2156, 2024.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

Canalys. Canalys Newsroom - Market Analysis and Research. https://canalys.com/newsroom, 2025. Accessed: 2025-02-23.

Capozzoli, A. and Primiceri, G. Cooling systems in data centers: State of art and emerging technologies. *Energy Procedia*, 83:484–493, 2015. doi: 10.1016/j.egypro.2015.11.484.

Chen, H., Waheed, A., Li, X., Wang, Y., Wang, J., Raj, B., and Abdin, M. I. On the diversity of synthetic data and its impact on training large language models. *arXiv preprint arXiv:2410.15226*, 2024a.

Chen, J., Dong, Q., Cheng, Y., Mahoney, M. W., and Ramchandran, K. Fedfm: Federated fine-tuning of foundation models under system and statistical heterogeneity. *arXiv preprint arXiv:2402.02777*, 2024b.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.

Chung, H. W., Hou, L., et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022.

Cobbe, K., Kosaraju, V., Bavarian, M., et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

CounterPoint. How many smartphones are in the world?, 2021. URL https://www.counterpointresearch.com/insights/smartphone-storage-capacity-zooms-increased-demand/.

Databricks. Free dolly: Introducing the world's first truly open instruction-tuned llm. https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm, 2023.

DeepLearning.AI. Introduction to federated learning. https://www.deeplearning.ai/short-courses/intro-to-federated-learning/, 2024. Accessed: 2025-02-23.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, FL, USA*, 2009.

Dettmers, T., Pagnoni, A., Holtzman, A., and Zettlemoyer, L. Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314*, 2023.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2019.

Dohmatob, E., Feng, Y., Subramonian, A., and Kempe, J. Strong model collapse. *arXiv preprint arXiv:2410.04840*, 2024.

Dong, X., Fu, Y., Diao, S., Byeon, W., Chen, Z., Mahabaleshwarkar, A. S., Liu, S.-Y., Bilicki, M. V. K., Ma, Z., Ai, Q., et al. Hymba: A hybrid-head architecture for small language models. *arXiv preprint arXiv:2411.13676*, 2024.

Du, Y. and Wei, F. Chinese tinyllm: Pre-training and fine-tuning of low-parameter language models for chinese. *arXiv preprint arXiv:2404.05480*, 2024.

Fan, L., Chen, K., Krishnan, D., Katabi, D., Isola, P., and Tian, Y. Scaling laws of synthetic images for model training. *arXiv preprint arXiv:2306.09387*, 2023a.

Fan, T., Kang, Y., Ma, G., Chen, W., Wei, W., Fan, L., and Yang, Q. Fate-llm: An industrial grade federated learning framework for large language models. *arXiv preprint arXiv:2310.10049*, 2023b.

Fang, X., Che, S., Mao, M., Zhang, H., Zhao, M., and Zhao, X. Bias of ai-generated content: an examination of news produced by large language models. *Scientific Reports*, 14(1):5224, 2024.

FedML-AI. Fedml: The unified and scalable ml library for large-scale distributed training, model serving, and federated learning. https://github.com/FedML-AI/FedML, 2024.

Fedus, W., Zoph, B., and Shazeer, N. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

Feng, S., Niyato, D., Wang, P., Kim, D. I., and Liang, Y.-C. Learning with quality guarantee in mobile edge computing systems. *IEEE Transactions on Mobile Computing*, 18(11):2506–2520, 2019.

FLock. Flock: Federated machine learning on the blockchain. https://www.flock.io/, 2023.

Flowerlab. Flower: A friendly federated ai framework, 2025. URL https://flower.ai/.

Frantar, E., Ashkboos, S., Hoefler, T., et al. Sparsegpt: Massive language models can be accurately pruned in one-shot. *arXiv preprint arXiv:2301.00774*, 2023.

Fu, Y., Peng, H., Khotilovich, A., Chen, L., and Yang, Y. Specializing smaller language models towards multi-step reasoning. *arXiv preprint arXiv:2301.12726*, 2023.

Gao, L., Biderman, S., Black, S., et al. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.

Glorioso, D., Zuo, S., and Jiang, D. Zamba2-2.7b: A state-of-the-art small language model achieving twice the speed and 27% reduced memory overhead. *arXiv preprint arXiv:2401.08642*, 2024.

Google Team. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024a.

Google Team. Gemma 2: Introducing the next generation of gemma open models. 2024b.

Grand View Research. Edge computing market size & share analysis report, 2023-2030, 2023. URL https://www.grandviewresearch.com/horizon/outlook/edge-computing-market-size/global.

Groeneveld, D., Beltagy, I., Davis, P., et al. Olmo: Accelerating the science of language models. *arXiv preprint arXiv:2402.00838*, 2024.

Gu, A. and Dao, T. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

Gunasekar, S., Zhang, Y., et al. Textbooks are all you need. *arXiv preprint arXiv:2306.11644*, 2023.

Hardy-White, D. Are ai scaling laws hitting a wall? https://www.linkedin.com/pulse/ai-scaling-laws-hitting-wall-dean-hardy-white-xchfe/, 2024. Accessed: 2025-01-22.

Hobbhahn, M., Heim, L., and Aydos, G. Trends in machine learning hardware, 2023. URL https://epoch.ai/blog/trends-in-machine-learning-hardware. Accessed: 2025-01-27.

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D. d. L., Hendricks, L. A., Welbl, J., Clark, A., et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

Horvath, S., Laskaridis, S., Almeida, M., Leontiadis, I., Venieris, S., and Lane, N. Fjord: Fair and accurate federated learning under heterogeneous targets with ordered dropout. *Advances in Neural Information Processing Systems*, 34:12876–12889, 2021.

Hu, E., Huang, W., et al. Minicpm: Unveiling the potential of small language models. *arXiv preprint arXiv:2402.03216*, 2024.

Inan, H., Khattab, O., Ainslie, J., Baevski, A., Blecher, T., Bui, C., Chowdhery, A., Dai, Z., Dong, L., Dyer, C., et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Javaheripi, M., Lobo, J., et al. Phi-2: The surprising power of small language models. *arXiv preprint arXiv:2312.12397*, 2023.

Jiao, X., Yin, Y., Shang, L., Jiang, X., Chen, X., Li, L., Wang, F., and Liu, Q. Tinybert: Distilling bert for natural language understanding. *arXiv preprint arXiv:1909.10351*, 2020.

Jin, L., Zhang, Y., Li, Y., Wang, S., Yang, H. H., Wu, J., and Zhang, M. Moe^2: Optimizing collaborative inference for edge large language models. *arXiv preprint arXiv:2501.09410*, 2025. URL https://arxiv.org/abs/2501.09410. Submitted to IEEE/ACM Transactions on Networking.

Jouppi, N. P., Young, C., Patil, N., and Patterson, D. In-datacenter performance analysis of a tensor processing unit. In *Proceedings of the 44th annual international symposium on computer architecture*, pp. 1–12, 2017.

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., et al. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210, 2021.

Kang, J., Xiong, Z., Niyato, D., Ye, H., and Kim, D. I. Incentive design for efficient federated learning in mobile networks: A contract theory approach. In *IEEE VTS Asia Pacific Wireless Communications Symposium*, pp. 1–5, 2019.

Kang, Y., Hauswald, J., Gao, C., Rovinski, A. M., Mudge, T., Mars, J., and Tang, L. Neurosurgeon: Collaborative intelligence between the cloud and mobile edge. In *Proceedings of the Twenty-Second International Conference on Architectural Support for Programming Languages and Operating Systems*, pp. 615–629. ACM, 2017. doi: 10.1145/3037697.3037698.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Khan, L. U., Pandey, S. R., Tran, N. H., Saad, W., Han, Z., Nguyen, M. N. H., and Hong, C. S. Federated learning for edge networks: Resource optimization and incentive mechanism. *IEEE Communications Magazine*, 57(10): 94–100, 2019.

Kim, S., Hooper, C., Gholami, A., et al. Squeezellm: Dense-and-sparse quantization. *arXiv preprint arXiv:2306.07629*, 2023.

Kressel, H. The end of moore's law? innovation in computer systems continues. *Artificial Intelligence in Science: Challenges, Opportunities and the Future of Research*, 2023. doi: 10.1787/a8d820bd-en.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 25, pp. 1097–1105, 2012.

Labs, E. Exo: Run your own ai cluster at home with everyday devices. https://github.com/exo-explore/exo, 2025. Accessed: 2025-01-29.

Lacoste, A., Luccioni, A., Schmidt, V., and Dandres, T. Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700*, 2019.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., and Soricut, R. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*, 2020.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998a.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998b.

Lee, K., Ippolito, D., Nystrom, A., Zhang, C., Eck, D., Callison-Burch, C., and Carlini, N. Deduplicating training data makes language models better. *arXiv preprint arXiv:2107.06499*, 2021.

Li, E., Zhou, Z., and Chen, X. Edge intelligence: On-demand deep learning model co-inference with device-edge synergy. In *Proceedings of the 2018 ACM/IEEE Symposium on Edge Computing*, pp. 31–46. IEEE, 2018a. doi: 10.1109/SEC.2018.00008.

Li, N. et al. Scaling data-constrained language models. *arXiv preprint arXiv:2403.03915*, 2024a.

Li, R., Choi, D., Chung, J., et al. Starcoder: May the source be with you! *arXiv preprint arXiv:2305.06161*, 2023.

Li, T., Sahu, A. K., Zaheer, M., Sanjabi, M., Talwalkar, A., and Smith, V. Federated optimization in heterogeneous networks. *arXiv preprint arXiv:1812.06127*, 2018b.

Li, T., Sanjabi, M., Beirami, A., and Smith, V. Fair resource allocation in federated learning. In *International Conference on Learning Representations*, 2020.

Li, W., Zhang, X., Wang, Y., and Chen, H. Domain-specific large language models: A comprehensive survey. *ACM Computing Surveys*, 57(2):1–38, 2024b.

Li, X., Wang, Y., and Chen, H. Distributed training of large language models: Challenges and opportunities in democratizing ai. *Computing Surveys*, 56(1):1–35, 2024c.

Li, X. et al. Carbon footprint reduction for sustainable data centers in real-time, 2024d.

Li, Y., Ding, Z., Ding, R., Ding, Y., Ding, X., and Ding, Y. Fedpet: Federated parameter-efficient fine-tuning of large language models. *arXiv preprint arXiv:2401.11597*, 2024e.

Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024a.

Liu, B., Bubeck, S., Eldan, R., Kulkarni, J., Li, Y., Nguyen, A., Ward, R., and Zhang, Y. Tinygsm: Achieving 80% on gsm8k with small models. *arXiv preprint arXiv:2312.09237*, 2023.

Liu, J., Chen, Y., Liang, Z., et al. Mobilellm: Empowering mobile large language models via architectural co-design. *arXiv preprint arXiv:2402.14905*, 2024b.

Longpre, S., Hou, L., Vu, T., et al. The flan collection: Designing data and methods for effective instruction tuning. *arXiv preprint arXiv:2301.13688*, 2023.

Lu, Z., Li, X., Cai, D., Yi, R., Liu, F., Zhang, X., Lane, N. D., and Xu, M. Small language models: Survey, measurements, and insights. *arXiv preprint arXiv:2409.15790*, 2024.

Ma, S., Zhao, H., Xue, L., et al. The era of 1-bit llms: All large language models are in 1.58 bits. *arXiv preprint arXiv:2402.17764*, 2024.

Magnusson, K., Sutskever, I., and Brown, T. B. Paloma: A benchmark for evaluating language model performance across tasks. *arXiv preprint arXiv:2306.07841*, 2023.

McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. Communication-efficient learning of deep networks from decentralized data. *Artificial Intelligence and Statistics*, pp. 1273–1282, 2017a.

McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. Communication-efficient learning of deep networks from decentralized data. *Artificial intelligence and statistics*, pp. 1273–1282, 2017b.

Mehta, H., Bartoldson, B., Shen, J., et al. Openelm: An efficient language foundation model. *arXiv preprint arXiv:2404.17766*, 2024.

Men, Y., Zhang, X., Sun, R., et al. Shortgpt: Layers in large language models are more redundant than you expect. *arXiv preprint arXiv:2402.18952*, 2024.

Meta. Introducing llama 3.1: Our most capable models to date. https://ai.meta.com/blog/meta-llama-3-1/, 2024. Accessed: 2025-01-22.

Meta AI. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Mitra, A., Mukherjee, S., et al. Orca 2: Teaching small language models how to reason. *arXiv preprint arXiv:2311.11045*, 2023.

Mohri, M., Sivek, G., and Suresh, A. T. Agnostic federated learning. In *International Conference on Machine Learning*, pp. 4615–4625, 2019.

Mukherjee, S., Mitra, A., Jawahar, G., Ordonez, V., and Chang, K.-W. Orca 2: Teaching small language models how to reason. *arXiv preprint arXiv:2312.02558*, 2023.

Muralidharan, S. et al. Minitron: Efficient heterogeneous transformer via pruning and distillation. *arXiv preprint arXiv:2407.07900*, 2024.

NanoReview.net. NanoReview.net - Gadget Specifications and Comparisons. https://nanoreview.net, 2025. Accessed: 2025-02-23.

Nayak, S., Patgiri, R., Waikhom, L., and Ahmed, A. A review on edge analytics: Issues, challenges, opportunities, promises, future directions, and applications. *Digital Communications and Networks*, 10(3):783–804, 2024.

NVIDIA. Nvidia jetson agx orin tflops specifications, 2023. URL https://forums.developer.nvidia.com/t/whether-tflops-on-tensor-cores-is-sparse-tflops-or-dense-tflops/248630. Forum discussion clarifying sparse vs. dense TFLOPS.

OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023a.

OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023b.

Parliament, E. and of the European Union, C. General data protection regulation (gdpr). 2016.

Patterson, D., Gonzalez, J., Le, Q. V., Liang, C., Munguia, L.-M., Rothchild, D., So, D., Texier, M., and Dean, J. Carbon emissions and large neural network training. *arXiv preprint arXiv:2104.10350*, 2021.

Penedo, G., Crnisanin, A., Shen, E., et al. The refinedweb dataset for falcon llm: Outperforming curated corpora with web data. *arXiv preprint arXiv:2306.01116*, 2023.

Peng, B., Alcaide, E., Anthony, Q., Albalak, A., Arcadinho, P., Cao, E., Cui, X., Dai, Z., Eissman, J., Firat, O., Fu, S., Gao, C., Hu, Y., Hughes, M., Kenealy, J., Krikun, M., Li, S., Li, Y., Liu, X., Luo, L., McAllester, D., Olson, M., Patel, Reiner Pope, A., Rao, N., Roberts, A., Shazeer, N., Siddhant, A., Tay, Y., Tran, D., Wang, J., and Wei, W. Rwkv: Reinventing rnns for the transformer era. *arXiv preprint arXiv:2305.13048*, 2023.

Pfeiffer, J., Dinu, S., et al. H2o-danube3: Raising the bar in open llms. *arXiv preprint arXiv:2407.21587*, 2024.

PrimeIntellect-ai. Opendiloco: An open-source framework for globally distributed low-communication training, 2025. URL https://github.com/PrimeIntellect-ai/OpenDiloco.

Qiu, X., Parcollet, T., Beutel, D. J., Topal, T., Mathur, A., and Lane, N. D. Can federated learning save the planet? *arXiv preprint arXiv:2010.06537*, 2021.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21: 1–67, 2020.

Raina, R., Madhavan, A., and Ng, A. Y. Large-scale deep unsupervised learning using graphics processors. In *Proceedings of the 26th annual international conference on machine learning*, pp. 873–880, 2009.

Rosenblatt, F. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

Sani, L., Iacob, A., Cao, Z., Lee, R., Marino, B., Gao, Y., Cai, D., Li, Z., Zhao, W., Qiu, X., et al. Photon: Federated llm pre-training. *arXiv preprint arXiv:2411.02908*, 2024.

ScaleFlux Research. Ai's hardware hunger: The global semiconductor supply chain under pressure. *ScaleFlux Insights*, 2024. URL https://scaleflux.com/in-the-media/ai-hardware-hunger-the-global-semiconductor-supply-chain-under-pressure/. Accessed: 2025-01-27.

Schwartz, R., Dodge, J., Smith, N. A., and Etzioni, O. Green ai. *Communications of the ACM*, 63(12):54–63, 2020.

Seagate. Dataage white paper: The digitization of the world – from edge to core, 2019. URL https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf.

Seagate. Rethink data report 2020, 2020. URL https://www.seagate.com/content/dam/seagate/migrated-assets/www-content/our-story/rethink-data/files/Rethink_Data_Report_2020.pdf.

Sevilla, J., Heim, L., Ho, A., Besiroglu, T., Hobbhahn, M., and Villalobos, P. Compute trends across three eras of machine learning. *2022 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–10, 2022.

Sharir, O., Peleg, B., and Shoham, Y. The cost of training nlp models: A concise overview. *arXiv preprint arXiv:2004.08900*, 2020.

Shridhar, K., Liu, X., Hasan, M., and Rashid, M. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. *arXiv preprint arXiv:2305.02301*, 2023.

Shumailov, I., Shumaylov, Z., Zhao, Y., Gal, Y., Papernot, N., and Anderson, R. The curse of recursion: Training on generated data makes models forget. *arXiv preprint arXiv:2305.17493*, 2023.

Smith, R. Nvidia announces jetson tx2: Parker comes to nvidia's embedded system kit. *IEEE Hot Chips*, 29, 2017.

Staff, B. Apple, nvidia secure future with taiwan semi's advanced chips as ai demand soars. *Benzinga*, June 2024. URL https://www.benzinga.com/news/24/06/39310821/.

Statista global data volume. Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2025, 2023. URL https://www.statista.com/statistics/871513/worldwide-data-created/.

Statista IoT device data volume. Internet of things (iot) connected devices data size worldwide from 2019 to 2025, 2023. URL https://www.statista.com/statistics/1017863/worldwide-iot-connected-devices-data-size/.

Sun, Y., Chen, Y., Li, Z., Liu, Z., Han, J., Ma, H., Wu, Y., and Liu, Q. Revisiting the scaling laws of neural language models. *arXiv preprint arXiv:1712.09444*, 2017.

Sun, Z., Chen, C., Zhang, Z., et al. A simple and effective pruning approach for large language models. *arXiv preprint arXiv:2306.11695*, 2023.

Systems, C. and Computer, T. Slimpajama: A 627b token cleaned and deduplicated version of redpajama. *arXiv preprint arXiv:2305.12661*, 2023.

Thawakar, O., Vayani, A., Khan, S., Cholakal, H., Anwer, R. M., Felsberg, M., Baldwin, T., Xing, E. P., and Khan, F. S. Mobillama: Towards accurate and lightweight fully transparent gpt. *arXiv preprint arXiv:2402.16840*, 2024.

Thompson, J., Askell, A., and Song, J. Frontier ai regulation: Managing emerging risks to public safety. *arXiv preprint arXiv:2207.05257*, 2022.

Thompson, N. C., Greenewald, K., Lee, K., and Manso, G. F. The computational limits of deep learning. *arXiv preprint arXiv:2007.05558*, 10, 2020.

Thompson, N. C., Greenewald, K., Lee, K., and Manso, G. F. Deep learning's diminishing returns. *IEEE Spectrum*, 58 (10):50–55, 2021.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Van Nguyen, C., Shen, X., Aponte, R., Xia, Y., Basu, S., Hu, Z., Chen, J., Parmar, M., Kunapuli, S., Barrow, J., et al. A survey of small language models. *arXiv preprint arXiv:2410.20011*, 2024.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Velasevic, B., Parasnis, R., Brinton, C. G., and Azizan, N. On the effects of data heterogeneity on the convergence rates of distributed linear system solvers. In *2023 62nd IEEE Conference on Decision and Control (CDC)*, pp. 8394–8399. IEEE, 2023.

Villalobos, P. and Ho, A. Trends in training dataset sizes. *Epoch AI Blog*, 2022. URL https://epochai.org/blog/trends-in-training-dataset-sizes.

Villalobos, P., Ho, A., Sevilla, J., Besiroglu, T., Heim, L., and Hobbhahn, M. Position: Will we run out of data? limits of llm scaling based on human-generated data. In *Forty-first International Conference on Machine Learning*.

Villalobos, P., Ho, A., Sevilla, J., Besiroglu, T., Heim, L., and Hobbhahn, M. Will we run out of data? limits of llm scaling based on human-generated data. *arXiv preprint arXiv:2211.04325*, pp. 13–29, 2024.

Wang, F., Zhang, Z., Zhang, X., Wu, Z., Mo, T., Lu, Q., Wang, W., Li, R., Xu, J., Tang, X., et al. A comprehensive survey of small language models in the era of large language models: Techniques, enhancements, applications, collaboration with llms, and trustworthiness. *arXiv preprint arXiv:2411.03350*, 2024a.

Wang, H., Yurochkin, M., Sun, Y., Papailiopoulos, D., and Khazaeni, Y. Optimizing federated learning on non-IID data with reinforcement learning. *IEEE International Conference on Computer Communications*, pp. 1698–1707, 2020.

Wang, H., Ma, S., Dong, L., Huang, S., Wang, H., Ma, L., Yang, F., Wang, R., Wu, Y., and Wei, F. Bitnet: Scaling 1-bit transformers for large language models. *arXiv preprint arXiv:2310.11453*, 2023.

Wang, H., Kaplan, Z., Lipton, Z. C., and Xing, E. P. Fedqa: Federated learning of large language models for question answering. *arXiv preprint arXiv:2402.03882*, 2024b.

Wang, Y., Li, X., Zhang, W., and Chen, H. Edullm: Tailoring large language models for personalized education. *Learning and Education*, 5(2):78–92, 2024c.

Wang, Y. E., Wei, G.-Y., and Brooks, D. Benchmarking tpu, gpu, and cpu platforms for deep learning. *arXiv preprint arXiv:1907.10701*, 2019.

Wenger, E. Ai produces gibberish when trained on too much ai-generated data, 2024.

Wireless, C. Edge Computing for IoT, Real-Time Data and Low Latency Processing, 2023. URL https://www.cavliwireless.com/blog/nerdiest-of-things/edge-computing-for-iot-real-time-data-and-low-latency-processing.html. Accessed:2025-01-22.

Wolf, T. et al. Rene: A hybrid model architecture for efficient language understanding. *arXiv preprint arXiv:2405.07236*, 2024.

Wu, J., Zhang, M., and Yang, H. H. Collaborative edge computing for democratized ai: A comprehensive review. *IEEE Internet of Things Journal*, 11(2):1123–1142, 2024a.

Wu, M., He, H., Wang, C., Wang, Z., Wang, Y., Xu, G., et al. Lamini-lm: A diverse herd of distilled models from llama-2. *arXiv preprint arXiv:2304.14402*, 2024b.

Wu, Z., Zhao, X., Wang, Z., Li, X., and Yu, P. S. Fedllm: Federated fine-tuning of large language models. *arXiv preprint arXiv:2402.03764*, 2024c.

xAI. Introducing grok-3. https://x.ai/blog/grok-3, 2025. Accessed: 2025-02-23.

XenonStack. Edge ai for personalization. https://www.xenonstack.com/blog/edge-ai-for-personalization, 2023. Accessed:2025-01-22.

Xu, J., Wang, Z., Yang, C., Li, T., and Liu, Y. Fedfm: Bridging the gap between federated learning and foundation models. *arXiv preprint arXiv:2401.11619*, 2024.

Yang, J., Wang, S., Ma, S., Zheng, J., et al. Qwen2: Technical report. *arXiv preprint arXiv:2404.05169*, 2024a.

Yang, L., Xu, M., Zhao, Z., Zhang, Y., and Chen, H. Fingpt: Large language models in quantitative finance. *Journal of Finance*, 79(1):145–178, 2024b.

Yang, Q., Liu, Y., Chen, T., and Tong, Y. Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(2):1–19, 2019.

Yang, Y. et al. Environmental burden of united states data centers in the artificial intelligence era, 2024c.

Yao, Y. et al. The pursuit of fairness in artificial intelligence models: A survey, 2024.

Ye, R., Ge, R., Zhu, X., Chai, J., Yaxin, D., Liu, Y., Wang, Y., and Chen, S. Fedllm-bench: Realistic benchmarks for federated learning of large language models. *Advances in Neural Information Processing Systems*, 37:111106–111130, 2025.

Ye, S., Du, J., Zeng, L., Ou, W., Chu, X., Lu, Y., and Chen, X. Galaxy: A resource-efficient collaborative edge ai system for in-situ transformer inference. *arXiv preprint arXiv:2405.17245*, 2024. URL https://arxiv.org/abs/2405.17245.

Yi, C., Zhai, P., Mathur, V., Singh, P., Basu, S., Hosseini, H., and Xu, R. Phonelm: A small language model with embedded phone system app capabilities. *arXiv preprint arXiv:2405.08131*, 2024.

Zhan, Y., Li, P., Qu, Z., Zeng, D., and Dou, S. Learning to incentivize: A bilateral neural network framework for efficient federated learning. *arXiv preprint arXiv:2004.03700*, 2020.

Zhang, J., Gu, J., Liu, S., et al. Quantized training of gradient scale-invariant neural networks. *arXiv preprint arXiv:2212.11552*, 2022.

Zhang, K., Chen, Z., et al. Towards making the most of chatgpt for machine translation. *arXiv preprint arXiv:2309.02654*, 2023a.

Zhang, P., Chen, G., et al. Tinyllama: An open-source small language model. *arXiv preprint arXiv:2401.02385*, 2024a.

Zhang, W., Chen, X., Liu, Y., and Wang, J. Democratizing ai: A survey of open-source initiatives and community-driven development in large language models. *arXiv preprint arXiv:2401.08562*, 2024b.

Zhang, W., Lu, Z., and Wang, B. Blockchain-based incentive mechanisms for federated learning. *IEEE Transactions on Services Computing*, 2024c.

Zhang, Y., Duan, Q., Guo, S., et al. Loraprune: Pruning meets low-rank parameter-efficient fine-tuning. *arXiv preprint arXiv:2305.18403*, 2023b.

Zhang, Y., Liu, W., Chen, J., Wang, Y., and Li, X. Medicalgpt: Domain adaptation of large language models for medical applications. *Nature Medicine*, 30(2):312–324, 2024d.

Zhao, Z., Gan, L., Wang, G., Hu, Y., Shen, T., Yang, H., Kuang, K., and Wu, F. Retrieval-augmented mixture of lora experts for uploadable machine learning. *arXiv preprint arXiv:2406.16989*, 2024.

Zhong, H., Wang, C., Levy, O., Choi, Y., and Smith, N. A. Legalbench: A framework for evaluating legal large language models. *arXiv preprint arXiv:2403.07985*, 2024.

A. Impact Statements

The shift from centralized to distributed training of large models, may introduce new technical and societal challenges and have the potential to fundamentally reshape the AI landscape.

A.1. AI Monopoly and Democratization

The current AI landscape is characterized by significant concentration of power among a few tech giants, primarily due to their monopoly over massive computing resources and data centers (Bommasani et al., 2021). This monopolistic trend has intensified with companies like OpenAI increasingly moving towards closed systems. While open-source alternatives like Llama (Touvron et al., 2023), Deepseek (Liu et al., 2024a) and other community-driven models have made strides towards democratization by releasing model parameters and technical reports (Zhang et al., 2024b), the gap in computational resources and data access between major AI companies and other players remains substantial and continues to widen. This disparity in resources allows tech giants to maintain their absolute dominance in determining the direction of AI development, raising concerns about AI democratization.

Edge device-based collaborative training presents a promising pathway to democratize AI development (Wu et al., 2024a). By leveraging the collective computing power of millions of edge devices, this approach could effectively challenge the existing monopolistic structure (Anderson et al., 2024b; Li et al., 2024c). This democratization of AI training through edge devices could fundamentally reshape the structure of responsibilities and authorities. If everyone can participate in training LLMs, the AI landscape could fundamentally change. Training decisions would shift from companies to communities, creating shared responsibility for model development (Brown et al., 2024). Global participation would help models reflect diverse cultural perspectives, while allowing communities to adapt models for their local needs. Furthermore, this decentralized approach could foster a more competitive and innovative AI ecosystem. When the barriers to entry for AI model training are lowered, we can expect to see a broader range of specialized models emerging (like (Li et al., 2024b; Zhang et al., 2024d; Yang et al., 2024b; Zhong et al., 2024; Anderson et al., 2024a; Wang et al., 2024c)), better suited to local needs and diverse use cases.

A.2. Fairness and Incentive Mechanisms

The distributed training paradigm introduces new considerations for model fairness and bias mitigation (Yao et al., 2024; Li et al., 2020). When training occurs across diverse edge devices, the resulting models can potentially better reflect the heterogeneous nature of user populations (Wang et al., 2020). However, this approach also raises concerns about participation bias, where differences in device capabilities or user engagement could lead to underrepresentation of certain groups (Kairouz et al., 2021). To address these challenges, researchers have proposed various fairness-aware federated learning algorithms (Mohri et al., 2019) that aim to ensure equitable model performance across different demographic groups and device types (Li et al., 2020).

To sustain a distributed training ecosystem, effective incentive mechanisms are crucial for motivating user participation (Kang et al., 2019). Traditional approaches like computational resource sharing (Khan et al., 2019) and privacy-preserving reward systems (Zhan et al., 2020) have shown promise in encouraging user engagement. More innovative solutions include token-based reward systems (Zhang et al., 2024c) and reputation mechanisms (Yang et al., 2019) that compensate users for their contributions while maintaining system integrity. These incentive structures not only encourage consistent participation but also help ensure the quality of contributed training data (Feng et al., 2019), creating a sustainable ecosystem for collaborative AI development.

A.3. Carbon Footprint and Energy Efficiency

The shift from centralized to distributed training offers compelling environmental benefits (Yang et al., 2024c). Traditional data centers housing large language models face significant energy challenges(Li et al., 2024d) - their high-performance GPUs require extensive cooling systems that consume 30-40% of total energy (Capozzoli & Primiceri, 2015). In contrast, FL distributes computation across edge devices like smartphones and tablets that operate at much lower temperatures and power levels, eliminating industrial cooling needs (Qiu et al., 2021).

FL also dramatically reduces data transmission energy costs. While centralized approaches require raw data transfer from millions of devices, FL only transmits lightweight model updates, substantially decreasing network energy overhead (McMahan et al., 2017a). The hardware efficiency gap is striking - edge devices like the NVIDIA Tegra X2 consume just 7.5W during training compared to 250W for data center GPUs (Smith, 2017), translating to major carbon footprint reductions, particularly for simpler models (Li et al., 2018b). By reducing reliance on power-hungry data centers and leveraging existing consumer devices, FL enables more sustainable AI development through optimized energy efficiency and minimized infrastructure needs. This combination of reduced cooling requirements, efficient hardware utilization, and optimized data handling makes FL an environmentally responsible choice for the future of AI training (Lacoste et al., 2019). As climate impact becomes increasingly critical, FL's sustainability advantages position it as a key technology for green AI development.

B. Historical Development and Current Challenges

B.1. Data: the fuel of LLMs

Early data-driven AI development As LLMs continue to achieve unprecedented success in artificial intelligence, understanding the role of data becomes increasingly crucial. From the early days of simple datasets to the modern era of massive data collections, data has consistently served as the lifeblood of AI, determining the upper bounds of model capabilities. The evolution of AI—marked by breakthroughs in computer vision, natural language processing, and beyond—can be traced back to the continuous expansion and refinement of data resources.

In the early stages of AI, despite relatively small data scales, the importance of data was already evident. The MNIST dataset, for instance, serves as a notable example. With 60,000 training images and 10,000 test images, it provided a crucial foundation for neural network research, demonstrating the fundamental role of data in model training (LeCun et al., 1998b). As data scales expanded, the capabilities of deep learning models saw significant improvements. The emergence of ImageNet, which contains 14 million images across 21,000 synsets, revolutionized computer vision. This enabled deep learning models like AlexNet to learn complex visual features and achieve breakthrough progress in image recognition tasks, reducing error rates from 26.2% to 15.3% in the ILSVRC-2012 competition (Deng et al., 2009; Krizhevsky et al., 2012). ImageNet's success stemmed not only from its scale but also from its high quality and diversity, laying the groundwork for subsequent large-scale data applications.

Era of massive data With the proliferation of the internet and advances in computing power, data scales have expanded dramatically, ushering AI into an era of massive data. GPT-3, for instance, was trained on 450 billion tokens, with a carefully curated mix of data sources: Common Crawl (60%), books (16%), Wikipedia (3%), and other internet-based text (21%) (Brown et al., 2020). This massive dataset enabled GPT-3 to excel across various tasks, demonstrating the decisive role of data scale in model capabilities. Compared to early datasets like MNIST and ImageNet, GPT-3's data scale and quality reached unprecedented heights, not only advancing natural language processing but also opening new possibilities for AI generalization.

Quality and diversity matter Beyond scale, data quality and diversity are crucial factors in model performance. ImageNet ensures data quality through rigorous validation, with each image verified by an average of 3.3 annotators and achieving 95% accuracy in its labels (Deng et al., 2009). This precise annotation enables models to learn accurate visual features and excel in image classification tasks. In the realm of large language models, GPT-3's training data underwent stringent cleaning and filtering, including deduplication, quality scoring based on document length and linguistic complexity, and content filtering for inappropriate content (Brown et al., 2020). This high-quality data enables GPT-3 to generate coherent and accurate text. Furthermore, diversity is essential: ImageNet covers 1,000 object categories across various domains, while GPT-3's training data spans multiple languages, genres, and knowledge domains, providing rich linguistic knowledge and contextual understanding.

Data as the ceiling for model capabilities A model's capability depends on the knowledge it extracts from data, following empirically observed scaling laws. While increasing model parameters can enhance expressive power, without sufficient data, models cannot effectively utilize these parameters. DeepMind's research on the Chinchilla model demonstrated that under the same compute budget, a 70B parameter model trained on 1.4T tokens outperforms a 280B parameter model trained on 0.35T tokens, achieving a 30% reduction in loss while using the same compute resources (Hoffmann et al., 2022). This finding directly supports the notion that data acts as a ceiling for model capabilities. Additionally, Meta's research shows that while Llama 2 (70B) has 70 billion parameters, its performance largely benefits from training on 2T tokens of high-quality data, with particular emphasis on academic papers, code repositories, and books that enhance its reasoning capabilities (Touvron et al., 2023). These studies emphasize data's central role in model training and suggest that optimal model scaling requires a balanced increase in both parameters and training data.

Looking ahead From MNIST to ImageNet to GPT-3, advances in data scale, quality, and diversity have directly driven AI breakthroughs. Data remains the foundation of AI development, determining the upper limits of model capabilities. As we push the boundaries of LLM performance, the challenge of acquiring sufficient high-quality, diverse data becomes increasingly acute. Traditional data sources like the internet are showing signs of exhaustion, and concerns about data privacy and ownership are growing. This motivates the exploration of novel data acquisition approaches, such as leveraging edge devices and distributed data collection, which we will explore in subsequent sections. The future of LLMs may depend not just on scaling existing data sources, but on fundamentally rethinking how we collect, curate, and utilize data in AI training.

B.2. Computing power: the engine of LLMs

Early neural networks and CPU era Since the inception of neural networks, every breakthrough in the field of AI has been driven by the continuous improvement of computational power (Thompson et al., 2020). From the early multilayer perceptron (MLP) to the widely used large language models (LLM) today, the progress in computing power has always been a key engine for advancing AI.

As the prototype of neural networks, the MLP was initially used to solve linearly separable problems (Rosenblatt, 1958). Due to its relatively low computational demand, it could run on traditional CPU environments. However, as the complexity of neural network models increased and application scenarios expanded, computational requirements gradually rose. The emergence of Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) marked a surge in computational demands. CNN, through convolutional operations, effectively reduced the number of parameters, enhancing the computational efficiency of image processing tasks. Classic models such as LeNet (LeCun et al., 1998a) and AlexNet (Krizhevsky et al., 2012) achieved significant results in image classification, but this also led to a surge in computational resource demands. For example, AlexNet's victory in the 2012 ImageNet competition was made possible by using the NVIDIA GTX 580 GPU, which significantly boosted computational performance (Krizhevsky et al., 2012).

GPU and TPU revolution With the growing scale of neural network models, GPUs gradually became indispensable computing tools (Raina et al., 2009). The parallel computing capabilities of GPUs greatly accelerated the training process of neural networks, particularly in the field of deep learning. Meanwhile, specialized hardware for deep learning, such as Tensor Processing Units (TPUs), emerged (Jouppi et al., 2017). Compared to GPUs, TPUs offer higher efficiency and lower power consumption when performing matrix operations and deep learning tasks (Wang et al., 2019), making them the preferred hardware for training large-scale neural networks.

Transformer era and computational demands As computational resources continued to expand, the scale of neural network model training also grew. The introduction of the Transformer architecture (Vaswani et al., 2017) revolutionized the field of natural language processing (NLP), especially with the launch of models like BERT (Devlin et al., 2018) and the GPT series (Brown et al., 2020; OpenAI, 2023b), which pushed NLP technology to new heights. However, the self-attention mechanism in the Transformer architecture has a computational complexity of $O(n^2)$, where n represents the sequence length (Vaswani et al., 2017). This means that as the model scale and sequence length increase, the required computational power grows exponentially. For example, training large language models like GPT-3 (Brown et al., 2020) and GPT-4 (OpenAI, 2023b) involves trillions of parameters and requires thousands of GPUs or TPU nodes to support the process. This immense computational demand not only places extremely high requirements on hardware, but also on computational frameworks, storage, and communication bandwidth, creating unprecedented challenges (Patterson et al., 2021).

Computing power as the key driver Every leap in Artificial Intelligence has been driven by computational power (Amodei & Hernandez, 2018). From multilayer perceptrons to convolutional neural networks, and the introduction of the Transformer architecture, every innovation in models has been accompanied by an explosive growth in computational needs (Thompson et al., 2020). Particularly in the era of large language models, computational power is not only the foundational tool for model training but also the core driving force behind breakthroughs in AI performance (Kaplan et al., 2020). The success of large-scale models like GPT-4 validates that AI progress almost entirely depends on the support of more powerful computational resources (Hoffmann et al., 2022).

C. Smartphone Data Volume Estimation

In the absence of publicly available, granular data on per-user smartphone data generation patterns, we adopt a conservative estimation approach to approximate the total annual smartphone data volume. While this method necessarily involves simplifications, it provides a robust lower-bound approximation that is sufficient to support our core arguments without compromising the validity of our conclusions.

Data volume estimation per smartphone: Based on industry reports (CounterPoint, 2021), the average smartphone storage capacity reached 100 GB in 2020. To ensure a conservative estimate, we assume that only 1% of this storage capacity (equivalent to approximately 1 GB per smartphone) is actively used for data generation and storage, including local images, video information, and other types of user-generated content. This assumption aligns with baseline usage scenarios while intentionally underestimating actual data utilization.

Number of smartphones: The growth of the number of smartphone users is an important basis for estimating the total amount of data. For this, we have referred to data from market research institutions (BankMyCell, 2023), which includes trends in changes to the number of smartphone users over time.

Based on the above statistical data, the total annual smartphone data volume D_{total} is calculated using the following formula:

$$D_{\text{total}}(\text{EB}) = N_{\text{users}} \times 1\,\text{GB/user} \times 10^{-3}\,(\text{conversion from GB to EB}), \tag{1}$$

where N_{users} represents the global smartphone user base in billions.

Substituting $N_{\text{users}} = 8.0 \times 10^9$ (representing 8 billion users) into Equation (1):

$$D_{\text{total}} = 8.0\,\text{GB/user} \times 10^{-3} = 8.0\,\text{EB}.$$

Our purpose is to establish a defensible lower bound for analysis. Even under these stringent assumptions, the derived volumes remain orders of magnitude higher than synthetic or centralized datasets, thereby reinforcing the strategic importance and value of edge-generated data. This conservative estimation underscores the critical need for scalable solutions capable of managing and leveraging such vast quantities of distributed data effectively.

D. Estimation of Smartphone Total Computational Power

To assess the (ideally) aggregate computational capabilities of smartphones globally, we estimate the total computing power, given the current lack of comprehensive statistical data in this domain. Our approach leverages two key data sources: the annual worldwide shipment volumes for major smartphone brands, and the computational performance specifications of mobile processors deployed in their devices during each corresponding year. The complete data underlying our analysis is presented in Table 1, which provides a detailed breakdown by manufacturer and time period. For quantitative analysis, we formulated a mathematical model to calculate the total computing power. Specifically, for any given year, we compute the aggregate computational capacity (C_{total}) by summing the contributions from each smartphone manufacturer (i). Each manufacturer's contribution is determined by multiplying their total device shipments (N_i) by the average computing power of their mobile processors (P_i) for that year, expressed formally as:

$$C_{\text{total}} = \sum_i N_i \cdot P_i \tag{2}$$

This formulation enables us to systematically track the evolution of distributed computing power across the smartphone ecosystem while accounting for both market share dynamics and technological advancement in mobile processors. By maintaining conservative estimates for processor capabilities and focusing on verified shipment data, our analysis provides a reliable lower bound for the total computational resources available through smartphones.

E. Small Language Model (SLM) Architectures and Training Methods

Table 2 presents a comprehensive overview of the Small Language Model (SLM) landscape, categorized by architectures and training methodologies, according to Wang et al. (2024a). The table is organized into two main categories: (I) Transformer-Based Models, which represent the dominant architecture in current SLMs, and (II) Alternative Architecture Models, which explore novel approaches to achieve efficiency. The Transformer-Based section is further divided into models pre-trained from scratch, models derived from larger LLMs through knowledge distillation, and models created through various compression techniques (pruning, quantization, etc.). The Alternative Architecture section showcases emerging approaches like State Space Models (Mamba, Hymba), recurrent architectures (RWKV, xLSTM), and traditional encoder-decoder or encoder-only designs.

This classification showcases the architectural innovations and training methodologies that are driving the SLM field forward, providing essential technical foundations for deploying powerful AI capabilities on resource-constrained edge devices. By documenting various model sizes, training corpora, and development techniques, the table offers a comprehensive overview of cutting-edge approaches that enable sophisticated language processing directly on end-user devices. These advancements represent critical building blocks for the next generation of on-device AI systems that can operate efficiently without constant cloud connectivity while still delivering robust performance across diverse applications.

Table 1. Trends in Smartphone Shipments and Compute Power. (Data source: (Canalys, 2025)).

Company	Shipments (Million units)	Chip Performance Range (TFLOPS)	Total Compute Power Contribution (EFLOPS)
2020			
Samsung (20%)	255.5	1.20–1.53	349
Apple (16%)	207.2	0.65	135
Xiaomi (12%)	149.6	0.24–1.20	108
OPPO (9%)	119.4	0.24–1.20	86
vivo (9%)	112.6	0.24–1.20	81
Others (33%)	420.5	0.04–0.24	59
Overall: Shipments = 1265 Million, Compute Power = 817 EFLOPS			
2021			
Samsung (20%)	274.5	1.42–1.72	430
Apple (17%)	230.1	1.71–1.94	420
Xiaomi (14%)	191.2	0.82–1.74	240
OPPO (11%)	145.1	0.82–1.74	180
vivo (10%)	129.9	0.82–1.74	160
Others (28%)	379.4	0.27–0.82	207
Overall: Shipments = 1350 Million, Compute Power = 1637 EFLOPS			
2022			
Samsung (22%)	257.9	0.49–2.01	322
Apple (19%)	232.2	1.79	416
Xiaomi (13%)	152.7	1.01–3.49	351
OPPO (10%)	113.4	1.01–3.49	261
Transsion (6%)	73.1	0.24–0.98	44.6
Others (31%)	364.1	0.84–1.31	393
Overall: Shipments = 1193 Million, Compute Power = 1788 EFLOPS			
2023			
Apple (20%)	229.1	2.15	493
Samsung (20%)	225.5	2.01–2.77	539
Xiaomi (13%)	146.1	2.15–3.99	449
OPPO (9%)	100.7	2.15–3.99	309
Transsion (8%)	92.6	0.24–1.31	72
Others (30%)	347.9	0.24–2.15	416
Overall: Shipments = 1142 Million, Compute Power = 2278 EFLOPS			
2024			
Apple (18%)	225.9	1.91–2.29	474
Samsung (18%)	222.9	3.38–3.41	758
Xiaomi (14%)	168.6	3.38–4.95	703
Transsion (9%)	106.7	0.05–0.67	38
OPPO (8%)	103.6	3.38–4.95	432
Others (33%)	395.4	0.05–1.72	352
Overall: Shipments = 1223 Million, Compute Power = 2758 EFLOPS			

F. Distributed Collaborative Frameworks

Distributed collaborative frameworks enable the deployment, training, and fine-tuning of language models across multiple devices or servers. Table 3 presents a comparison of prominent frameworks in this domain. These frameworks can be broadly categorized into three types: cloud-based platforms that offer centralized resources for distributed computing, federated learning systems that enable training across decentralized data sources while preserving privacy, and fully decentralized frameworks that distribute computation across peer nodes. Some frameworks like Neurosurgeon (Kang et al., 2017), MoE^2 (Jin et al., 2025), Edgent (Li et al., 2018a), and Galaxy (Ye et al., 2024) focus on collaborative inference by partitioning models between edge devices and servers. Others, such as FedLLM (Wu et al., 2024c), FedFM (Chen et al., 2024b), FedPET (Li et al., 2024e), and Photon (Sani et al., 2024), specialize in federated fine-tuning of large language models while maintaining data privacy. These frameworks are essential for enabling efficient deployment of language models in resource-constrained environments and for scenarios requiring privacy preservation or operation in disconnected settings.

Table 2. Small Language Model (SLM) Architectures and Training Methods

Model	Sizes	Architecture	From Scratch	From LLMs	Training Method	Datasets
I. Transformer-Based Models						
I.A. Pre-Trained from Scratch						
PhoneLM (Yi et al., 2024)	0.5B; 1.5B	Transformer	✓		Pre-training	DCLM-baseline (Li et al., 2024a), StarCoderData (Li et al., 2023)
Llama 3.2 (AI, 2024)	1B; 3B	Transformer	✓		Pre-training, SFT, RLHF, DPO	Not released (9T tokens)
Qwen 1/1.5/2/2.5 (Yang et al., 2024a; Bai et al., 2023)	0.5B-7B	Transformer	✓		Pre-training	Not released
Gemma/Gemma 2 (Google Team, 2024a;b)	2B; 7B	Transformer	✓		Pre-training	Unknown
SmolLM (Allal et al., 2024)	135M-1.7B	Transformer	✓		Pre-training	SmolLM corpus (Benallal et al., 2024)
H2O-Danube3 (Pfeiffer et al., 2024)	500M; 4B	Transformer	✓		Pre-training (multi-stage)	Unknown
MiniCPM (Hu et al., 2024)	1.2B; 2.4B	Transformer	✓		Pre-training	Dolma (AI2, 2023), C4 (Raffel et al., 2020)
CT-LLM (Du & Wei, 2024)	2B	Transformer	✓		Pre-training	MAP-CC
OLMo (Groeneveld et al., 2024)	1B; 7B	Transformer	✓		Pre-training	Dolma (AI2, 2023) (multiple sources)
TinyLlama (Zhang et al., 2024a)	1B	Transformer	✓		Pre-training	SlimPajama (Systems & Computer, 2023)
Phi-series (Abdin et al., 2024; Javaheripi et al., 2023)	1.3B-6.6B	Transformer	✓		Pre-training	CodeTextBook (Gunasekar et al., 2023)
OpenELM (Mehta et al., 2024)	270M-3B	Transformer	✓		Pre-training	RefinedWeb (Penedo et al., 2023), PILE (Gao et al., 2020)
MobiLlama (Thawakar et al., 2024)	0.5B; 0.8B	Transformer	✓		Pre-training	LLM360 Amber
MobileLLM (Liu et al., 2024b)	125M; 350M	Transformer	✓		Pre-training	Unknown (1T tokens)
I.B. Derived from Larger Models						
MINITRON (Muralidharan et al., 2024)	4B	Transformer		✓	Distillation, Pruning	8T tokens from Nemotron-4
Orca/Orca 2 (Mitra et al., 2023; Mukherjee et al., 2023)	7B; 13B	Transformer		✓	Distillation	Orca 2 dataset, FLAN-v2 (Longpre et al., 2023)
MINIMA (Zhang et al., 2023a)	3B	Transformer		✓	Distillation (from Llama-2-7B)	Pile (Gao et al., 2020), Wudao
Dolly-v2 (Databricks, 2023)	3B; 7B	Transformer		✓	Instruction tuning (from Pythia)	Databricks-dolly-15k
LaMini-LM (Wu et al., 2024b)	61M-7B	Transformer		✓	Distillation	LaMini instruction dataset
I.C. Model Compression Approaches						
SparseGPT (Frantar et al., 2023)	Various	Transformer		✓	Unstructured Pruning	Not applicable
Wanda (Sun et al., 2023)	Various	Transformer		✓	Unstructured Pruning	Not applicable
LoRAPrune (Zhang et al., 2023b)	Various	Transformer		✓	Unstructured Pruning	Not applicable
ShortGPT (Men et al., 2024)	Various	Transformer		✓	Structured Pruning	Not applicable
BitNet/BitNet b1.58 (Wang et al., 2023; Ma et al., 2024)	Various	Transformer		✓	Quantization (QAT)	Not applicable
QLoRA (Dettmers et al., 2023)	Various	Transformer		✓	Quantization, Low-Rank	Various fine-tuning datasets
SqueezeLLM (Kim et al., 2023)	Various	Transformer		✓	Quantization (PTQ)	Not applicable
II. Alternative Architecture Models						
Mamba (Gu & Dao, 2023)	125M-1.3B	Mamba	✓		Pre-training	Pile (Gao et al., 2020)
Rene (Wolf et al., 2024)	1.3B	Mamba	✓		Pre-training	Dolma-1.7 (AI2, 2023)
Zamba2 (Glorioso et al., 2024)	2.7B	Mamba	✓		Pre-training	Not specified
Hymba (Dong et al., 2024)	125M-1.5B	Hymba	✓		Pre-training	DCLM-Baseline (Li et al., 2024a)
xLSTM (Beck et al., 2024)	125M-1.3B	xLSTM	✓		Pre-training	SlimPajama (Systems & Computer, 2023)
RWKV (Peng et al., 2023)	169M-14B	RNN	✓		Pre-training	Pile (Gao et al., 2020)
Specialized FlanT5 (Fu et al., 2023)	250M-3B	Encoder-Decoder		✓	Instruction Tuning	GSM8K (Cobbe et al., 2021)
FlanT5 (Chung et al., 2022)	80M-3B	Encoder-Decoder		✓	Instruction Tuning	Muffin, T0-SF, SNI and CoT
T5 (Raffel et al., 2020)	60M-3B	Encoder-Decoder	✓		Pre-training	C4 (Raffel et al., 2020)
DistilBERT (Sanh et al., 2019)	66M	Encoder-only		✓	Distillation (from BERT)	Wikipedia, BookCorpus
TinyBERT (Jiao et al., 2020)	14.5M	Encoder-only		✓	Distillation (from BERT)	Wikipedia, BookCorpus
ALBERT (Lan et al., 2020)	12M-18M	Encoder-only	✓		Pre-training (parameter sharing)	Wikipedia, BookCorpus

	Distributed Capabilities						
Framework	**Inference**	**Training**	**Pretraining**	**Fine-tuning**	**Type**	**Privacy**	**License**
exo-explore/exo (Labs, 2025)	✓				Decentralized		MIT
Together AI (AI, 2023)	✓	✓	✓	✓	Cloud		Commercial
FLock Platform (FLock, 2023)		✓		✓	Federated, Blockchain	✓	Apache 2.0
OpenDiloco (PrimeIntellect-ai, 2025)		✓	✓		Decentralized		Apache 2.0
FederatedScope (alibaba, 2024)		✓		✓	Federated	✓	Apache 2.0
FedML (FedML-AI, 2024)		✓		✓	Federated	✓	Apache 2.0
Flower (Flowerlab, 2025)		✓		✓	Federated	✓	Apache 2.0
FATE-LLM (Fan et al., 2023b)		✓		✓	Federated	✓	Apache 2.0
FedLLM (Ye et al., 2025)		✓	✓	✓	Federated	✓	CC BY-NC 4.0

Table 3. Comparison of Distributed Machine Learning Frameworks