

Beyond Efficiency: A Systematic Survey of Resource-Efficient Large Language Models

Guangji Bai¹, Zheng Chai², Chen Ling¹, Shiyu Wang¹,
Jiaying Lu¹, Nan Zhang³, Tingwei Shi¹, Ziyang Yu¹,
Mengdan Zhu¹, Yifei Zhang¹, Xinyuan Song¹, Carl Yang¹,
Yue Cheng², Liang Zhao^{1*}

¹*Department of Computer Science, Emory University, 201 Dowman Dr,
Atlanta, 30322, GA, United States.

²School of Data Science and Department of Computer Science,
University of Virginia, 1827 University Avenue, Charlottesville, 22904,
VA, United States.

³College of Information Sciences and Technology, Pennsylvania State
University, 201 Old Main, University Park, 16802, PA, United States.

*Corresponding author(s). E-mail(s): liang.zhao@emory.edu;
Contributing authors: guangji.bai@emory.edu; dub6yh@virginia.edu;
chen.ling@emory.edu; shiyu.wang@emory.edu; jiaying.lu@emory.edu;
njz5124@psu.edu; tshi30@emory.edu; zyu31@emory.edu;
mengdan.zhu@emory.edu; yifei.zhang2@emory.edu; xsong30@emory.edu;
j.carlyang@emory.edu; mrz7dp@virginia.edu;

Abstract

The burgeoning field of Large Language Models (LLMs), exemplified by sophisticated models like OpenAI’s ChatGPT, represents a significant advancement in artificial intelligence. These models, however, bring forth substantial challenges in high consumption of computational, memory, energy, and financial resources, especially in environments with limited resource capabilities. This survey aims to systematically address these challenges by reviewing a broad spectrum of techniques designed to enhance the resource efficiency of LLMs. We categorize methods based on their optimization focus—covering computational, memory, energy, financial, and network resources—and their applicability across various stages of an LLM’s lifecycle, including architecture design, pre-training,

fine-tuning, and system design. Additionally, the survey introduces a nuanced categorization of resource efficiency techniques by their specific resource types, which uncovers the intricate relationships and mappings between various resources and corresponding optimization techniques. A standardized set of evaluation metrics and datasets is also presented to facilitate consistent and fair comparisons across different models and techniques. By offering a comprehensive overview of the current state-of-the-art and identifying open research avenues, this survey serves as a foundational reference for researchers and practitioners, aiding them in developing more sustainable and efficient LLMs in a rapidly evolving landscape. To make the field more accessible to interested beginners, we not only make a systematic review of existing works and a highly structured taxonomy of resource-efficient LLMs but also release a website including a constantly-updated paper list <https://github.com/tiingweii-shii/Awesome-Resource-Efficient-LLM-Papers>.

Keywords: Large Language Model, Resource Efficiency, Sustainable AI, Survey.

1 Introduction

In recent years, Large Language Models (LLMs) [1, 2] have achieved significant advancements, redefining the frontier of artificial intelligence. These models, such as OpenAI’s GPT-3 with its impressive 175 billion parameters, represent a quantum leap in complexity and capability [3]. The trend in LLM development is toward ever-increasing model sizes, with some recent entrants boasting upwards of hundreds of billions of parameters [4–6]. This scale amplifies their utility across a spectrum of applications, from intelligent chatbots to intricate data analyses, and even facilitating research in diverse domains. However, the exponential growth in model sizes presents a huge demand for various resources (e.g., computation, energy, memory) [7–9]. The immense resource requirements to train or deploy such extensive models can be cost-prohibitive, particularly in resource-constrained environments like academic labs or the medical sector, which do not have access to the vast computational resources of major IT conglomerates. Additionally, the environmental impact is a growing concern, as the extensive GPU usage for training these models translates to significant electricity consumption and increasing carbon dioxide emissions [7]. Addressing these challenges requires a focused effort on enhancing the resource efficiency of LLMs at every stage of their lifecycle.

Defining *Resource-Efficient LLMs* requires an understanding of the critical resources involved in the lifecycle of LLMs. In this survey, we systematically categorize the essential resources into five key categories: *computation, memory, energy, money, and communication cost*. Computation refers to the processing power necessary to train and run these models; memory encompasses the data storage capacity required; energy denotes the electricity consumed during operation; financial resources pertain to the monetary investment needed for infrastructure and ongoing costs; and communication cost involves the bandwidth and latency in data transfer during training and inference. Efficiency in this context is characterized by the ratio of these

resources invested to the output produced, with a more efficient system being one that yields the same level of output while consuming fewer resources. A resource-efficient LLM, therefore, is designed to maximize performance and capabilities while minimizing resource expenditure across all these dimensions, thereby enabling more sustainable and accessible AI solutions.

Resource efficiency in LLMs is a crucial and complex area that demands innovative solutions to address significant challenges. These challenges, more pronounced in LLMs than in smaller neural networks like CNNs and MLPs, arise from the unique scale and complexity of LLMs. We outline these challenges from various key perspectives:

- [Model] 1. Low parallelism in auto-regressive generation: Auto-regressive token generation, the predominant method in LLMs, suffers from significant latency due to poor parallelism [10]. This is especially problematic for large model sizes or extensive input lengths, hindering efficient processing in both training and inference. 2. Quadratic complexity in self-attention layers: The multi-head self-attention layer in LLMs exhibits quadratic complexity with respect to the input sequence length [11]. This complexity creates a computational bottleneck as the input length increases, limiting the practical input sequence length LLMs can handle efficiently.
- [Theory] 1. Scaling laws and diminishing returns: theoretical insights into scaling laws for neural networks, particularly LLMs, suggest that as models become larger, the benefits in performance improvement per parameter added diminish [12]. This phenomenon raises questions about the optimal size of LLMs and the balance between resource investment and performance gain. 2. Generalization and overfitting: Theoretical work on generalization in machine learning is particularly relevant for LLMs [13, 14]. Understanding the limits of what large models can generalize from training data and the risks of overfitting is crucial for developing more resource-efficient models.
- [System] Given the substantial model size of LLMs and the vast training datasets, fitting them into the memory of a single GPU/TPU is unfeasible [15, 16]. Consequently, intricate system designs become crucial to optimize the training process for LLMs and successfully accomplish the task. Furthermore, the system design gains increased significance due to the latency and throughput requirements associated with the inference tasks of LLMs, particularly when taking into account user experience and the constraints of a limited cost budget [17, 18].
- [Ethics] 1. Dependence on large and proprietary training data: Many LLMs are trained on extensive, proprietary datasets, making it challenging to apply certain efficiency improvement techniques that require access to the original training data [19]. This limitation not only restricts the scope of potential improvements but also raises ethical questions about transparency and the democratization of AI advancements. 2. Closed source models and lack of parameter access: Many advanced LLMs are closed source [1, 20–22], with restricted access to their parameters. This constraint means that efforts to improve efficiency must be conducted without deep insights into the model’s internal workings, complicating the process of optimizing resource usage. The closed nature of these models also brings up ethical concerns regarding the concentration of AI capabilities and the openness of scientific research.

- **[Metrics]** In the context of LLMs, the development of comprehensive metrics for evaluating resource efficiency faces unique challenges due to the diverse and complex nature of LLM tasks and architectures [10]. Unlike smaller models where optimizing one or two resources, such as computation or memory, might be sufficient, LLMs present a multi-objective problem requiring simultaneous optimization across multiple key resources, including computation, memory, energy, monetary cost, etc [23]. Therefore, a comprehensive metric for LLMs must provide a holistic view that encapsulates all these critical resources, quantifying not only the individual resource usage but also the interdependencies and trade-offs between them [24]. This approach is crucial for advancing LLMs in a balanced and sustainable manner, significantly more complex than metric development for smaller models.

In recent years, significant research efforts have been dedicated to developing and applying resource-efficient LLMs to address the challenges referenced earlier. There has been a wave of research proposing and deploying new strategies across various fields, although the concept of resource-efficient LLMs is relatively nascent. Most existing LLM approaches have been tailored for specific application domains; however, the underlying principles are often adaptable enough to be utilized in other areas. Nevertheless, it remains challenging to compare these resource-efficient strategies across different domains that cater to distinct communities. Furthermore, assessing the performance of resource-efficient LLMs demands intricate and specialized evaluation strategies due to their distinctive attributes, such as their multi-dimensional efficiency (e.g., computational, energy, memory usage) and the diverse outcomes they produce. To date, there is a deficiency in systematic standardization and a comprehensive summarization framework to evaluate the various methodologies proposed for resource-efficient LLMs. This lack of a cohesive summary and classification of existing methods and applications in resource-efficient LLMs poses significant issues for practitioners who need clear information on current limitations, pitfalls, unresolved questions, and promising directions for future research.

In response to these gaps, this paper seeks to offer a systematic review of the techniques, benchmarks, and evaluation metrics that contribute to the resource efficiency of LLMs. To our knowledge, this constitutes the first detailed survey explicitly devoted to resource efficiency in the context of LLMs. In the following, we outline the principal contributions of this survey:

- **Comprehensive overview of resource-efficient LLM techniques:** Our paper makes a significant contribution by offering a comprehensive overview of techniques aimed at enhancing the resource efficiency of Large Language Models. This overview is extensive, covering the entire range of the LLM lifecycle. It delves into various methodologies and strategies developed in the field, focusing on how they contribute to making LLMs more resource-efficient.
- **Systematic categorization and taxonomy of techniques by resource type:** We established a systematic categorization and taxonomy of resource-efficient LLM techniques, organized primarily by the type of resource(s) they optimize. This taxonomy simplifies the process of identifying and selecting appropriate methods based on specific resource optimization needs and provides a clear

and organized framework that aids researchers and practitioners in navigating the landscape of resource-efficient LLMs.

- **Standardization of evaluation metrics and datasets:** We present a standardized set of evaluation metrics and datasets tailored for assessing the resource efficiency of LLMs. This standardization facilitates consistent and fair comparisons across different models and techniques and provides a benchmark for future research in the field.
- **Identification of gaps and future research directions:** The paper concludes with a thoughtful discussion of the current bottlenecks and unresolved challenges in creating resource-efficient LLMs. By examining the limitations of existing approaches, we shed light on potential avenues for future research.

1.1 Related work

In this section, we discuss the relationship between our survey and some existing surveys on similar topics. In general, we can divide existing surveys related to the efficiency and acceleration of LLMs into the following categories: 1. fundamental overview of LLMs; 2. survey of model compression for LLMs; and 3. review of techniques of efficiency and acceleration for general deep neural networks.

- **Fundamental overview of LLMs.** With the recent surge in the popularity and efficacy of LLMs, numerous review papers have surfaced, offering insights into various aspects of LLMs. Some concentrate on dissecting the fundamental components of LLMs [25–27], while others delve into the historical context and potential applications of generative AI [28–30]. A select few [31] explore strategies for enhancing LLMs with reasoning capabilities. Nevertheless, a comprehensive review and technical taxonomy specifically focused on the specialization of LLM domains remains an unaddressed gap in the current literature.
- **Survey of compression and acceleration for LLMs.** Transformer-based language models have achieved huge success, however, the computational and memory cost remains a big concern despite the superior performance. There have been several survey papers on how to compress and accelerate large language models. For example, some discuss how to accelerate the inference of LLMs [32–34], by focusing on model compression techniques. In addition, a select few [35, 36] explore more efficient and lightweight architecture designs for transformers, which are the backbone of modern LLMs. Furthermore, some works discuss the efficient training of LLMs [37]. However, those existing surveys either lack comprehensiveness or are not up-to-date, especially considering the large number of papers published after the birth of ChatGPT, which marks the beginning of the LLMs era.
- **Review of efficient deep neural networks.** How to achieve efficient design or accelerate the computation of deep neural networks (DNNs) has long been a popular research direction, and there have been a couple of survey papers on this topic. Some works focus on the model compression and acceleration of DNNs [38, 39]. A few others discuss the hardware design and optimization for DNNs [40, 41]. However, due to the very large model size and the special architecture of

transformers, there is a big gap in directly applying those techniques for DNNs onto the LLMs.

1.2 Outline

The remainder of this survey is structured as follows, offering a detailed exploration of resource-efficient LLMs:

- Section 2 *Preliminary and taxonomy*: This section sets the foundation by introducing the fundamental concepts behind transformers and pre-trained LLMs. It establishes a comprehensive taxonomy of resources essential for LLMs, such as computation, memory, energy, money, and network communication. This taxonomy serves as a guiding framework for the entire survey, outlining the key areas of focus for improving resource efficiency in LLMs.
- Section 3 *LLM architecture design*: This section delves into the latest developments in LLM architecture, emphasizing designs that enhance resource efficiency. It discusses both efficient transformer architectures, which optimize traditional transformer models, and non-transformer architectures which propose alternative structures for resource optimization.
- Section 4 *LLM pre-training*: This section explores the various pre-training techniques for LLMs, highlighting how they contribute to resource efficiency. Key areas such as memory efficiency, data efficiency, and innovative training pipeline designs are examined, illustrating how each technique impacts the overall resource utilization during the pre-training phase.
- Section 5 *LLM fine-tuning*: This section covers the fine-tuning phase of LLMs, focusing on methods that enhance resource efficiency. It includes detailed discussions on parameter-efficient fine-tuning, which minimizes parameter updates; and full-parameter fine-tuning, which optimizes the entire parameter set.
- Section 6 *LLM inference*: Here, we analyze various inference techniques that improve resource efficiency in LLMs. The section features discussions on static methods including pruning, quantization, knowledge distillation, low-rank approximation, etc. In addition, we also discuss dynamic methods such as dynamic inference, which adapts computational resources in real time, and token parallelism, which optimizes processing at the token level to enhance efficiency during the inference stage.
- Section 7 *System design*: This section addresses system-level strategies for resource-efficient LLMs, encompassing support infrastructure, which focuses on leveraging specialized systems for efficiency, and deployment optimization, which involves strategies for deploying LLMs in a resource-conscious manner.
- Section 8 *Technique categorization by resources*: In this section, we evaluate the effectiveness of various resource efficiency techniques. The discussion revolves around real-world applications and how different methods fare in practical scenarios, providing a bridge between theory and application.
- Section 9 *Benchmark and evaluation metrics*: This section presents the benchmarks and metrics used for evaluating the resource efficiency of LLMs. It highlights the importance of standardized evaluation criteria in assessing the effectiveness of various techniques and models.

- Section 10 *Open Challenges and future directions*: Here, we identify the existing challenges and potential future research directions in the field of resource-efficient LLMs. This section is crucial for understanding the current gaps in the field and where future efforts may be most beneficial.
- Section 11 *Conclusion*: The survey concludes with a summary of the key findings and insights presented, encapsulating the core takeaways from the exploration of resource efficiency in LLMs.

2 Preliminary and taxonomy

In this section, we first provide some preliminaries of this survey, including some introduction about transformers and LLMs. Then, we introduce our proposed taxonomy of the techniques for the efficiency and acceleration of LLMs.

2.1 Preliminaries

2.1.1 Transformer model

The Transformer model stands as a pivotal milestone in the evolution of deep learning, particularly in the realm of natural language processing (NLP). Introduced by [11], the Transformer model represents a groundbreaking departure from conventional sequence-to-sequence models, offering an innovative solution to the challenges of capturing long-range dependencies in sequences.

- *Embedding Layers*. The embedding layer is the foundational component of a Transformer model, serving as the initial step in transforming raw input data into a format that can be effectively processed. It maps discrete tokens, such as words or subwords, into continuous vector representations, often referred to as word embeddings. These embeddings capture semantic relationships between words and enable the model to understand the meaning of each token.
- *Positional Encoding*. Unlike recurrent neural networks (RNNs) or convolutional neural networks (CNNs), Transformers do not inherently possess knowledge of the order or position of tokens in a sequence. To address this limitation, positional encodings are introduced. These encodings are added to the word embeddings and provide the model with information about the position of each token within the sequence. Typically, sinusoidal functions are used to generate these positional encodings, ensuring that the model can capture sequential dependencies without relying on recurrence or convolution.
- *Self-Attention*. Self-attention is the cornerstone of the Transformer architecture and allows the model to weigh the importance of different words in the input sequence when making predictions for a particular word. It computes a weighted sum of all input words, where the weights are determined dynamically based on the similarity between words. The self-attention mechanism is computed using a weighted sum over all words in the sequence, and the weights are determined by

the dot product of query, key, and value vectors:

$$\text{Attention}(Q, K, V) = \text{softmax} \left(\frac{QK^T}{\sqrt{d_k}} \right) V. \quad (1)$$

Here, Q , K , and V are the query, key, and value matrices, respectively, and d_k is the dimension of the key vectors. This mechanism allows the model to focus on relevant information within the input sequence.

- *Multi-Head (Self-)Attention.* Multi-head self-attention extends the self-attention mechanism by performing it multiple times in parallel, with different sets of learned parameters. This allows the model to capture different types of relationships and dependencies in the input sequence, providing a richer representation. Mathematically, multi-head self-attention involves computing multiple sets of query, key, and value matrices, and then concatenating the results from each head:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \text{head}_2, \dots, \text{head}_h)W^O, \quad (2)$$

where $\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$ represents the output of the i -th attention head, and W^O is a learned linear transformation. Multi-head self-attention enhances the model's ability to capture both local and global dependencies in the data.

In summary, the Transformer architecture consists of these key components, each playing a crucial role in enabling the model to understand and generate sequences effectively, making it a powerful tool in natural language processing and other sequence-to-sequence tasks.

2.1.2 Large Language Models (LLMs)

Pre-trained Language Models (PLMs) constitute a type of neural network that has been trained on extensive collections of text data. Their purpose is to acquire knowledge of linguistic patterns, structures, and semantics inherent in the language. In the context of LLMs, the input comprises a text sequence that serves as the context for comprehension and processing. Often, a prompt or additional sentence is included to clarify the task. These prompts are tailored to the specific NLP task at hand, providing a premise or task explanation. For example, in text summarization, a prompt like “Summarize the key points in the following passage:” can be placed before the input passage. The output is the generated text sequence or prediction responding to the input, e.g., the summarized key points of the provided passage. In some cases, post-processing steps such as token decoding or label extraction may be necessary for the final presentation.

LLMs typically follow the architectural designs of PLMs and come in three primary flavors: encoder-only, encoder-decoder, and decoder-only architectures. Here's an overview of these LLM architectures and their distinctions:

- *Encoder-only Language Models.* These models process input text to create vector representations without an explicit decoding phase for generating new text.

Instead, they transform and embed text into a high-dimensional space. Encoder-only models are primarily designed to capture and understand patterns and semantics in the input data. They find extensive use in tasks such as text classification, sentiment analysis, and clustering. A notable example is BERT [42], which extracts context-rich embeddings for downstream tasks through pre-training on a masked language modeling objective.

- *Encoder-Decoder Language Models.* These models consist of an encoder that processes input text into vector representations and a decoder that generates output text based on these representations. They employ cross-entropy loss as the objective function, comparing the actual and predicted target sequences. Encoder-Decoder LLMs are often used for sequence-to-sequence tasks such as machine translation and summarization. T5 [43] is a notable example of this architecture.
- *Decoder-only Language Models.* Examples like GPT [44] are autoregressive language models that generate the next word in a sequence based on previous words. They map a sequence of tokens to a vector representation and generate contextually relevant content autoregressively, calculating the probability of the next token based on the context. This autoregressive approach is particularly suitable for text-generation tasks.

In summary, LLMs and their variants play a pivotal role in natural language processing tasks by leveraging pre-training on vast text corpora to facilitate a wide range of language understanding and generation tasks.

2.2 Proposed taxonomy

2.2.1 Taxonomy of key resources involved with using LLMs.

The taxonomy for resource efficiency in LLMs encompasses five key domains: computation, memory, energy, money, and network communication. Each domain addresses a distinct aspect of resource utilization:

- **Computation:** This involves the processing power required for tasks such as training, fine-tuning, and executing LLMs. Evaluating computational efficiency includes considering the number of operations (like floating-point operations), the efficiency of algorithms, and the utilization of processing units like GPUs or TPUs. It is crucial to explore how to maximize output while minimizing computational requirements.
- **Memory:** Memory efficiency pertains to the amount of RAM and storage needed. LLMs, especially those with billions of parameters, require significant memory for storing the model weights and for processing large datasets during training and inference. Optimizing data structures, employing techniques like model pruning, and exploring memory-efficient architectures are key strategies here.
- **Energy:** This resource refers to the electrical power consumed during the model's lifecycle. Given the environmental impact and operating costs, energy efficiency is vital. It includes strategies for reducing power consumption, such as optimizing hardware utilization, using energy-efficient hardware, and implementing algorithms that require less computational power.

- **Money:** Financial resources are a crucial consideration, especially for smaller organizations and researchers. This includes the cost of hardware acquisition, electricity for running the models, and potential cloud computing expenses. Finding ways to make LLMs accessible and viable for a broader range of users without significant financial investment is another key challenge.
- **Network communication:** For distributed training and cloud-based deployment, network bandwidth and latency become significant. Efficient network communication means reducing the amount of data that needs to be transferred between nodes in a distributed system or between the cloud and end-users, which can greatly affect training time and responsiveness in real-time applications.

2.2.2 Taxonomy of techniques for resource-efficient LLMs

As delineated in Figure 1, our survey paper introduces a structured taxonomy that categorizes techniques for enhancing the resource efficiency of LLMs into clear, defined tiers. We propose five principal categories: Architecture Design, Pre-training, Fine-tuning, Inference, and System Design. Each of these is selected for its integral role in the lifecycle of efficient LLM development and deployment.

- **Architecture design.** This category examines the structural foundations of LLMs, branching into Transformer-based and Non-transformer Architectures. These classifications are intended to highlight architectural variations and innovations crucial for the models' efficiency and efficacy.
- **Pre-training.** This category inspects the preliminary phases of LLM development, including Memory Efficiency and Data Efficiency. It underscores the importance of the pre-training environment and strategies that significantly affect the models' future performance and resource utilization.
- **Fine-tuning.** Addressing the optimization of pre-trained models, this category is organized into Parameter-efficient Fine-tuning and Full-parameter Fine-tuning. These subdivisions represent the range of techniques that refine models for particular tasks or enhance their overall functionality.
- **Inference.** During the operational stage, various strategies under the Inference category, such as Model Compression and Dynamic Acceleration, are employed. This classification acknowledges the diverse tactics applied at the model inference phase, impacting efficiency and performance distinctly.
- **System design.** Focusing on system-level considerations, this category covers Deployment Optimization and Support Infrastructure, among others. It explores hardware and system optimizations that are essential for improving the practical performance of LLMs.

Through this taxonomy, we aim to facilitate a structured and nuanced understanding of the diverse methodologies and strategies employed in the quest for enhanced efficiency and acceleration of LLMs, providing a holistic view of the current research landscape.

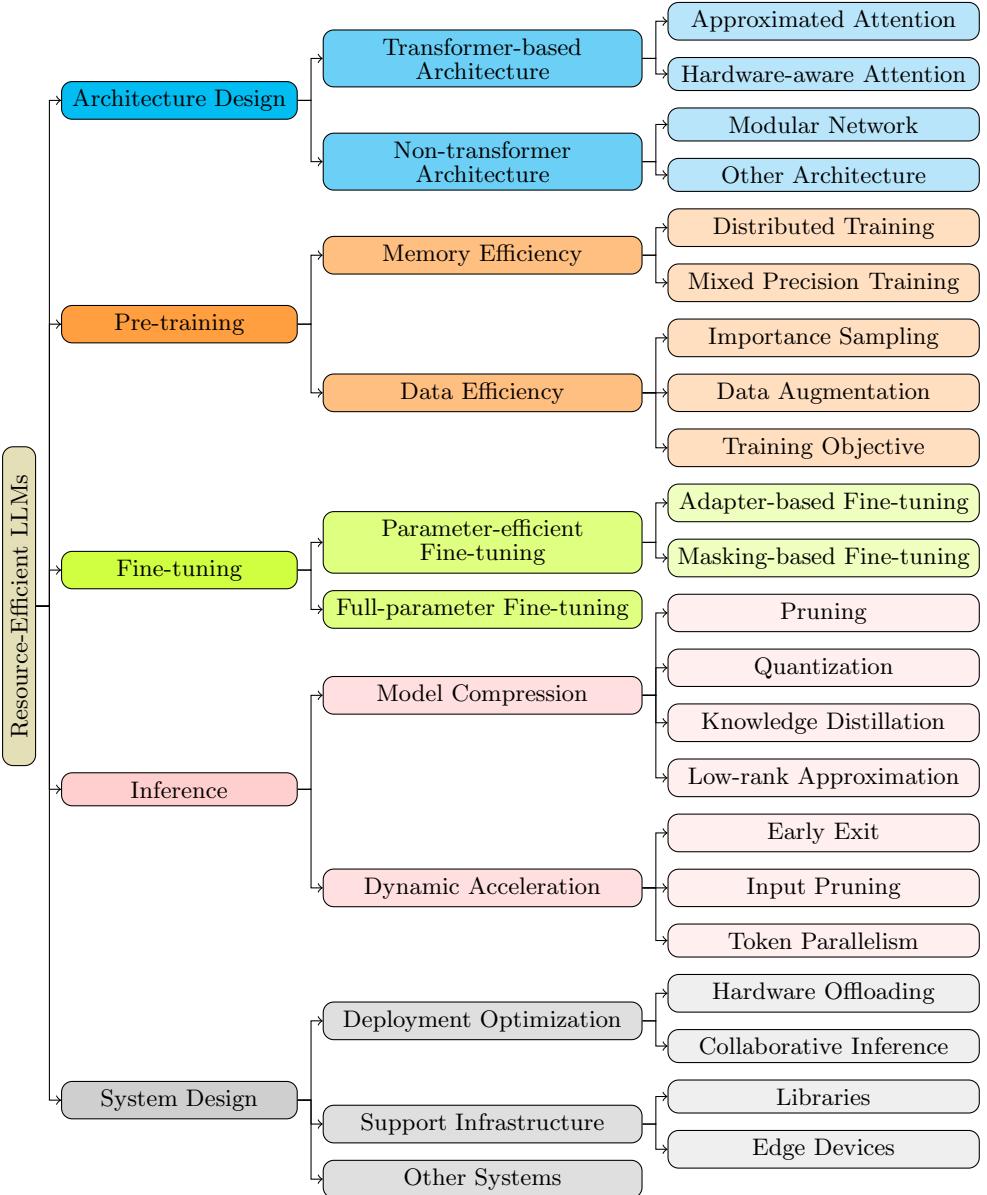


Fig. 1 A taxonomy of techniques for achieving resource-efficient LLMs.

3 LLM architecture design

This section explores the advancements in architecture design for LLMs, specifically focusing on enhancing the efficiency of Transformer models. We examine various strategies aimed at reducing computational and memory demands, crucial for the practical deployment of LLMs. The discussion includes innovative approaches like

Reformer, Linear Transformer, AFT, and KDEformer, each presenting unique solutions to optimize processing speed and resource usage. Additionally, we touch upon hardware-optimized attention mechanisms and alternative non-transformer architectures, highlighting their contributions to the evolving landscape of efficient LLM design.

3.1 Efficient transformer architecture

Efficient transformers focus on creating neural network architectures that are optimized for enhanced throughput. The attention layer significantly influences the processing speed of transformers, which contributes a lot to the throughput.

3.1.1 Approximate attention.

One stream of works focuses on designing attention operators with approximation techniques to achieve **less time complexity and/or less memory complexity**. In the classic Transformer, the time complexity of the self-attention operator is $\mathcal{O}(T^2d)$, and the memory complexity is $\mathcal{O}(T^2)$. Here T, d denote the sequence length and hidden feature dimension, respectively. Reformer [45] replace dot-product attention by proposed locality-sensitive hashing attention, which leads to $\mathcal{O}(Td \log T)$ time complexity and $\mathcal{O}(T \log T)$ memory complexity. Linear Transformer [46] expresses the self-attention as a linear dot-product of kernel feature maps and utilizes the associativity property of matrix products to reduce the complexity term from T^2 to T , thus achieving $\mathcal{O}(Td^2)$ time complexity and $\mathcal{O}(Td + d^2)$ memory complexity. EfficientAttention [47] proposes an approximated attention operation by switching the **QKV** multiplication order from $(\mathbf{QK}^\top)\mathbf{V}$ to $\mathbf{Q}(\mathbf{K}^\top\mathbf{V})$, which leads to more efficient $\mathcal{O}(T^2d)$ time complexity and same $\mathcal{O}(Td + d^2)$ memory complexity when $d < T$. AFT [48] proposes an extremely efficient variant called AFT-simple, which achieves linear complexity in both time ($\mathcal{O}(Td)$) and memory ($\mathcal{O}(Td)$). In an AFT layer, the key and value are first combined with a set of learned position biases ($s < T$), so that the multiplication between the key-value and query is in an element-wise manner. The introduced learned position bias s can be eliminated in the AFT-simple variant (*i.e.* no position bias is learned) so that AFT-simple completely gets rid of the need for dot products operations. Memory efficient attention [49] presents a practical implementation for self-attention that requires only $\mathcal{O}(d \log T)$ memory with the same time complexity $\mathcal{O}(T^2d)$. The core idea behind memory efficient attention is similar to “lazy softmax” [50] where the denominator of the softmax for the dot product of queries and keys can be calculated in the later stage. KDEFormer [51] suggests reducing the denominator of the softmax function to a variant of the kernel density estimation (KDE) problem, and an efficient KDE solver can be further utilized to accelerate the multiplication of the attention matrix with the value matrix. The trick is based on reducing the number of columns of the attention matrix (referred to as $A := \exp(QK^\top / \sqrt{d})$) using importance sampling. KDEFormer delivers a $\mathcal{O}(Tmd)$ time complexity and $\mathcal{O}(Tm)$ memory complexity, where $m < T$ is a small number. MEGA [52] introduces a moving average equipped gated attention mechanism to solve the weak inductive bias and quadratic computational complexity. MEGA offers

$\mathcal{O}(cTd)$ time complexity and $\mathcal{O}(cTd)$ memory complexity with a theoretical grounding, where $c < T$ is MEGA’s chunk size of quadratic attention. LoMA [53] introduces a method for losslessly compressing the memory of transformer-based language models, allowing for a substantial increase in contextual length without altering the model architecture. By segmenting inputs into reading, memory, and repetition areas, LoMA utilizes a bidirectional attention mask within the memory area to preserve information. The approach achieves up to a 4:1 compression ratio, maintaining the model’s generative capability through fine-tuning and enabling efficient long-text handling with minimal data requirements. BiPE [54] introduces a bilevel positional encoding approach that combines intra-segment and inter-segment encodings to enhance length extrapolation capabilities in transformer models. This design disentangles positional information within and between segments, allowing for more efficient encoding. BiPE achieves superior length extrapolation performance with a theoretical advantage, delivering a time complexity of $\mathcal{O}(Td)$ and memory complexity of $\mathcal{O}(Td)$ across diverse tasks. Simple Linear Attention [55] combines sliding window and linear attention mechanisms, offering a solution to the recall-throughput tradeoff by balancing memory consumption and token recall. It delivers $\mathcal{O}(Td^2)$ time complexity and utilizes a hardware-optimized IO-aware algorithm, achieving up to 24 \times higher throughput than FlashAttention-2, making it a highly efficient architecture for language generation. Cluster-wise Graph Transformer (Cluster-GT) introduces the Node-to-Cluster Attention (N2C-Attn) mechanism [56], leveraging Multiple Kernel Learning in a kernelized attention framework to capture node and cluster-level information without compressing clusters into single embeddings, achieving linear time complexity and excelling in graph-level tasks by integrating dual-granularity feature maps through an efficient cluster-wise message-passing architecture. SageAttention [57] introduces a novel quantization method specifically for the attention mechanism, achieving approximately 2.1 \times and 2.7 \times higher OPS than FlashAttention2 and xformers, respectively, while maintaining superior accuracy to FlashAttention3, thus enabling efficient model inference with minimal end-to-end performance loss across large language, image, and video generation models. Local Attention Mechanism (LAM) [58] leverages the continuity of time series data to reduce attention computations, achieving $\mathcal{O}(n \log n)$ time and memory complexity, significantly improving upon traditional $\mathcal{O}(n^2)$ complexity, and demonstrates superior performance in long-horizon forecasting, surpassing state-of-the-art models and addressing the need for new evaluation datasets in time series forecasting. The proposed Long LoRA Pereceiver (LLP) [59] framework builds upon the PerceiverAR architecture to effectively cut down the quadratic complexity of traditional Transformer-based attention, achieving semi-linear time complexity, and demonstrates notable improvements over existing state-of-the-art models, positioning LLP as a compelling and efficient core component for next-generation Large Language Models. Signformer [60] introduces a from-scratch transformer pipeline with novel convolution-attention integration, achieving substantial parametric (467-1807x) and computational efficiency over contemporary SOTAs, attaining near-LLM-level performance, securing the 2nd place on the leaderboard, and demonstrating the feasibility of sustainable, edge-deployable sign language translation without reliance on large pretrained models or extensive datasets

Approach	Time Complexity	Memory Complexity
Transformer [11]	$\mathcal{O}(T^2d)$	$\mathcal{O}(T^2 + Td)$
Reformer [45]	$\mathcal{O}(Td \log T)$	$\mathcal{O}(T \log T + Td)$
Linear Transformer [46]	$\mathcal{O}(Td^2)$	$\mathcal{O}(Td + d^2)$
Efficient Attention [47]	$\mathcal{O}(T^2d)$	$\mathcal{O}(Td + d^2)$
AFT [48]	$\mathcal{O}(Td)$	$\mathcal{O}(Td)$
Memory Efficient Attention [49]	$\mathcal{O}(T^2d)$	$\mathcal{O}(d \log T)$
KDEformer [51]	$\mathcal{O}(mTd)$	$\mathcal{O}(mT)$
MEGA [52]	$\mathcal{O}(cTd)$	$\mathcal{O}(cTd)$
Simple Linear Attention [55]	$\mathcal{O}(Td^2)$	$\mathcal{O}(Td^2)$
RWKV [61]	$\mathcal{O}(Td)$	$\mathcal{O}(d)$

Table 1 Overview of time complexity and memory complexity improvements for selected approaches over classical Transformer. Here, T, d denote the sequence length and hidden feature dimension, respectively. m used in KDEformer denotes m sampled columns of attention matrix. c used in MEGA denotes its chunk size of quadratic attention.

3.1.2 Hardware optimized attention.

There is another stream of works focusing on **hardware efficient attention operator**. Starting from 2021, many works (LightSeq [62], Faster Transformer [63], xFormers [64]) have been focused on optimizing CUDA implementation of attentions and transformer layers including kernels fusion, gemm optimization, *etc*. FlashAttention [65] introduces an IO-aware precise attention algorithm that employs tiling to minimize the volume of memory reads/writes between the high bandwidth memory of the GPU and the on-chip SRAM. Building on this, FlashAttention-2 [66] further refines FlashAttention by addressing the suboptimal work partitioning concern. vLLM [67] proposes a novel attention algorithm, PagedAttention, that mainly optimizes the virtual memory and paging techniques in operating systems. MobileLLM [68] introduces a deep-and-thin model structure optimized for on-device use cases, leveraging embedding sharing and grouped-query attention to enhance performance. With innovations such as block-wise weight sharing, MobileLLM achieves state-of-the-art results for sub-billion parameter models, offering $\mathcal{O}(Td)$ time complexity and maintaining model efficiency even in memory-constrained environments.

3.2 Non-transformer architecture

While Transformers, with their self-attention mechanisms, have dominated the field of language modeling, alternative architectures have emerged to tackle various challenges or provide different advantages.

3.2.1 Modular network

Modular Network (also called the Mixture of Experts (MoEs)) [69, 70] technique is a machine learning approach that combines multiple specialized models, known as experts, to solve complex tasks more effectively. As we know, a single dense LLM itself contains billions of parameters, which is extremely difficult to be further scaled into larger parameter sizes. MoE provides a solution to enabling LLM's parameter size

to grow from hundreds of billions into trillions, by sparse routing (also mentioned as *sparse activation*, *sparse gating*). During training, multiple individual expert LLMs and a routing function are trained simultaneously. The learned routing function allows the MoE system to select a subset of experts according to the input, thus reducing computational and memory requirements. Switch Transformer [71] follows the principle of “increasing the parameter count while keeping the floating point operations (FLOPs) per example constant”. Switch Transformer achieves this by replacing the original dense feed-forward network layer in the Transformer with a sparse Switch feed-forward layer. The Switch layer is essentially similar to [69] architecture but the authors simplify the number of selected experts to 1. GLaM [72] has 1.2T parameters which is 7X larger than GPT-3 and requires half of the FLOPs for inference. GLaM is implemented with 64 experts per MoE layer where each input token only activates 96.6B (8% of 1.2T) parameters. Different from the Switch Transformer, it contains an MoE layer interleaved with a traditional Transformer layer in each block of GLaM. Concurrent works include heterogeneous MoEs [73], MoE-LM [74], Unified Routing Network [75].

3.2.2 Other architecture

Researchers have explored more novel dense architectures that are different from the transformers. Inspired by AFT introduced in Section 3.1, RWKV [61] combines the efficient parallelizable training of Transformers with the efficient inference of RNNs. The key idea behind RWKV is to leverage a linear attention mechanism so that the proposed model can be formulated as a transformer during training and an RNN during inference. On the other hand, [76] explores the potential of Multi-Layer Perceptrons trained with the same next-token prediction to achieve non-trivial performance on text generation and arithmetic tasks. The authors also supply rich theory analysis to connect and compare their proposed architecture to existing transformer-based architectures, and they argue that the power of LLM can mostly be attributed to the large-scale auto-regressive next-token training scheme. Hyena [77] proposes a subquadratic drop-in replacement for attention constructed by interleaving implicitly parametrized long convolutions and data-controlled gating. Monarch Mixer [78] utilizes a simple yet efficient subquadratic *generalized matrix multiply algorithms* based architecture. Mamba [79] integrates *selective state space* models into a simple end-to-end neural network architecture without attention blocks, achieving competitive modeling power of Transformers while scaling linearly in sequence length. YOCO [80] introduces a memory-efficient decoder-decoder architecture that caches key-value pairs once, enabling sublinear scaling in GPU memory consumption and substantially reducing prefilling latency across long-context language models. MatMul-free LM [81] eliminates costly matrix multiplication operations by leveraging ternary weights and element-wise Hadamard products, achieving substantial reductions in memory usage and training latency while scaling to billion-parameter language models. RWKV-edge [82], including low-rank approximation, sparsity predictors, and clustering head, effectively reduce RWKV model size by 4.95–3.8x with a minimal 2.95pp drop in accuracy, facilitating the practical deployment of RNN-based LLMs on resource-constrained devices

and demonstrating the viability of efficient, high-performing large language models in embedded environments.

4 LLM pre-training

For large-scale LLMs like GPT-4, efficient pre-training is pivotal due to their extensive size and complexity. This efficiency transcends mere speed, focusing on optimal utilization of computational resources and innovative data management. Combining advanced hardware, such as GPUs and TPUs, with techniques like data and model parallelism, the pre-training process is tailored to be resource-efficient. Additionally, strategies like selective data sampling, model pruning, and quantization play a crucial role in minimizing data and memory requirements. These methods collectively contribute to not only accelerating the training process but also ensuring sustainable and cost-effective development of advanced LLMs.

4.1 Memory efficiency

4.1.1 Distributed training

Distributed training proves to be a highly effective approach for accelerating model training, particularly for machine learning tasks that exceed the memory capacity of a single accelerator, such as GPU, TPU, and more. In distributed training, the task of training the model is divided and allocated to multiple working nodes. These nodes concurrently execute local training tasks and collectively contribute to developing the original task.

Data parallelism. Data parallelism (DP) is the most straightforward approach for distributed training and has been inherently supported by famous machine learning frameworks like TensorFlow and PyTorch. In the paradigm of data parallelism, the initial dataset is divided into multiple partitions, and different data partitions are trained in parallel by multiple accelerators. However, DP has memory redundancies across all data partitions, and model states including model parameters, gradients, and optimizer states are required by each data partition. Given the substantial size of LLMs, applying DP to LLMs in a naive manner is impractical. To end this, ZeRO [15], PaLM [6] and Fairscale [83] introduce approaches for enhancing the efficiency of memory utilization in DP. Instead of duplicating the entire model states, these techniques suggest partitioning them. Each data partition stores a portion of the model states and can retrieve additional states from other data partitions with a dynamic communication schedule when necessary.

Model parallelism. Model parallelism (MP) is a kind of distributed training method that aims to minimize a model's memory footprint by spreading its layers or tensors across multiple accelerators, while DP primarily concerns data partitioning. Based on the partition levels, model parallelism can be categorized into two main types: tensor model parallelism (TMP) and pipeline model parallelism (PMP).

In the context of TMP, tensors can be split along their rows or columns, enabling concurrent execution of matrix multiplication operations across all split parts.

Megatron-lm [16] employs parallelization techniques for matrix multiplication operations within both the multi-layer perceptron (MLP) and the self-attention block of the transformer layer. In the case of the MLP, the weight matrix is divided along its columns, while for the self-attention block, the Query, Key, and Value parameters are split in a column-parallel fashion. Alternatively, Mesh-tensorflow [84] split the units in the hidden layer to achieve tensor MP.

In the case of PMP, a model is divided into multiple layer groups, and each accelerator is responsible for handling one of these groups. To minimize inter-accelerator communication, these groups typically consist of consecutive layers. While naively implementing PMP can reduce the memory demands on each accelerator, it is important to note that, due to layer dependencies, most accelerators are idle at any given time, with only one in active operation. To enhance the resource utilization, GPipe [85] and PipeDream [86] adopt an approach where a batch is divided into smaller micro-batches. PMP is then executed independently on each micro-batch, and gradient updates occur asynchronously across these micro-batches. BPipe [87] aims to achieve memory balance among accelerators during the training of PMP by transferring intermediate activations. Alpa [88] proposes a model-parallel training system for large deep-learning models. It has the capability to automatically generate parallel execution plans that encompass data, operator, and pipeline parallelisms. MegaScale [89] introduces a scalable training system leveraging 3D parallelism and in-depth observability, achieving high training efficiency and stability across over 10,000 GPUs. ProTrain [90] executes PMP independently on each micro-batch, allowing asynchronous gradient updates across these batches, thus enhancing throughput efficiency.

4.1.2 Mixed precision training

Mixed precision training is a technique used to accelerate the training of deep learning models by using both 16-bit and 32-bit floating-point types (as opposed to just using 32-bit or 64-bit throughout the training process such as BERT [42]). This approach has gained popularity, especially in the training of large language models, where computational cost can be a significant barrier. Recently, to pre-train extremely large language models, some works borrowed 16-bit floating-point numbers [91] which largely reduced memory consumption compared with 32-bit or 64-bit. To mitigate the performance degradation caused by the quantization with 16-bit floating-point numbers, Scao et al. [5] proposed Brain Floating Point (i.e., BF16) for training that is able to allocate more exponent bits and fewer significant bits than FP16.

4.2 Data efficiency

Data efficiency represents how efficiently a training pipeline leverages its data. It determines the number of iterations (steps) required to complete a training process, thus affecting the overall training cost. Since existing LLMs such as LLaMA [2] are usually trained on a large quantity of texts, maximizing the utilization of data offers a promising solution to reduce training cost.

Recent works try to improve data efficiency in various aspects of the training pipeline. We identify three major directions of achieving this goal: importance sampling, data augmentation, and training objective.

4.2.1 Importance sampling

A current survey [37] notes that importance sampling (data pruning) significantly influences models' data efficiency during pre-training. Importance sampling means to prioritize informative training instances, so it involves estimating per-sample importance. It is also called data pruning. A major solution of importance estimation is to compute gradient norm [92, 93]. More recent approaches [94, 95] work towards accelerating the data importance sampling process.

Data-Juicer [96], an LLM data processing system, enables efficient and scalable data processing to improve the quality of the training data. As a result, the generated data recipes from Data-Juicer yielded considerable improvements on LLaMA [2] performance in various pre-training and post-tuning cases. Similarly, INGENIOUS [97] is another system that aims to improve data quality by selecting highly representative subsets of the training corpora. ASTEROID [98], a multi-stage computational framework, first trains an MLFF (machine learning force fields) model on a large amount of inaccurate data to captures the sophisticated structures of training data and then fine-tunes the obtained model on a small amount of accurate data to improve model performance. Since inaccurate data is cheap while accurate data is much more expensive, ASTEROID improves data efficiency by fully utilizing the cheap inaccurate data. LISA [99], a memory-efficient fine-tuning method for LLMs, leverages layerwise importance sampling to selectively update model layers, thereby reducing GPU memory consumption while outperforming traditional full-parameter tuning and LoRA in downstream tasks.

4.2.2 Data augmentation

Data augmentation creates modified copies of existing data so that the current data can be fully utilized. Since it is an effective technique of improving data efficiency, a joint data augmentation for vision-language representation learning [100] is proposed to improve the existing pre-training pipelines. As an outcome of improving the data efficiency of pre-training pipelines, researchers of this work also show that downstream performance can be positively impacted. The training of generative adversarial networks (GANs) is also benefited by data augmentation [101]. Moreover, work has been done to augment acoustic data through pseudo acoustic representations of textual data to improve speech processing [102]. The proposed LLMRec framework enhances recommender systems by leveraging large language models to augment user-item interaction graphs, item attributes, and user profiles, thereby addressing data sparsity and improving recommendation accuracy [103]. A novel data augmentation technique, LLM-DA, leverages the text generation capabilities of large language models to enhance few-shot named entity recognition by generating semantically coherent and diverse training data [104].

4.2.3 Training objective

A recent survey [10] finds that the choice of pre-training objective is another factor that determines data efficiency. For the design of pre-training objective [105], it is typically a function of model architecture, input/target construction, and masking strategy. Specifically, representative masking strategies include masked language modeling [106], masked image modeling [107], and language-image pre-training [108]. Researchers of these works find that skipping the processing of some masked tokens can significantly improve training efficiency.

5 LLM fine-tuning

Fine-tuning Large Language Models (LLMs) like GPT-4 for specialized tasks involves a critical balance between achieving task-specific performance and maintaining resource efficiency, given their considerable size and computational demands. This section discusses various fine-tuning strategies, focusing on optimizing computational load, memory usage, and energy consumption. Techniques such as parameter-efficient fine-tuning, which adjusts a limited subset of parameters, offer a resource-conscious approach, while full-parameter fine-tuning, involving the modification of all parameters, is explored in the context of its higher resource requirements. This exploration is key to understanding how fine-tuning in LLMs is evolving to address the dual challenges of performance optimization and resource constraints.

5.1 Parameter-efficient fine-tuning

Parameter-efficient fine-tuning (PEFT) is a technique aimed at making the most out of an LLM’s vast parameter space without the need to adjust all the parameters during the fine-tuning process. Given the immense size of modern LLMs, fine-tuning every parameter can be computationally expensive and may even risk overfitting on smaller, task-specific datasets. Current PEFT techniques can be categorized into two main streams: 1) Masking-based Fine-tuning and 2) Adapter-based Fine-tuning.

Masking-based fine-tuning. In this approach, only a subset of the model’s parameters is updated during the fine-tuning process. The rest of the parameters are “masked” or frozen, meaning they are not updated during backpropagation. This masking could be applied to specific layers, certain types of parameters, or parameters identified through various criteria like importance scores. Previous research endeavors [109, 110] have focused on the optimization of fine-tuning procedures for comparatively smaller language models through the deployment of diverse regularization methodologies. However, these approaches exhibit limitations when applied to the fine-tuning of LLMs due to the significantly elevated computational requirements and voluminous datasets essential for the effective training of such models. In addressing these challenges, the method known as CHILD-TUNING [111] employs data from the target task to identify a subset of parameters—referred to as the “child network”—that are most pertinent to the task, while preserving the pre-trained values for parameters in the remaining architecture. In a related vein, a list of methods [112–114] introduces a dynamic parameter selection pipeline specifically tailored for the efficient fine-tuning of LLMs. These

works adaptively designate a judicious sub-network for staged updates, leveraging gradient information from back-propagation, thereby achieving notable performance enhancements on domain-specific tasks, particularly in resource-constrained environments. To optimize resource usage, MEFT [115] implements sparse activations and a Mixture of Experts (MoE) approach, dynamically offloading trainable parameters to the CPU and selectively transferring only the most relevant ones to the GPU, thereby reducing GPU memory load and communication overhead. To tackle the rank selection challenge, DyLoRA [116] leverages a dynamic low-rank adaptation approach, training LoRA blocks across a spectrum of ranks and enabling rank-specific inference without the need for costly search processes, thus optimizing efficiency while maintaining performance across a range of model sizes.

Adapter-based fine-tuning. Different from the previous method, in this approach, additional lightweight layers (adapters) are inserted between existing layers of the pre-trained model. During fine-tuning, only the parameters of these adapter layers are updated, while the original model parameters are kept fixed [117, 118]. Recent scholarly contributions have focused on Unsupervised Domain Adaptation (UDA) employing adapter mechanisms to advance the capabilities of pre-trained models in cross-lingual or multi-task learning contexts. A pioneering approach [119] involved multi-domain adaptation through a bifurcated strategy: an initial domain-fusion training phase employing Masked Language Model (MLM) loss on a composite corpus, followed by task-specific fine-tuning. A subsequent development, UDApter [120], extended this dual-phase methodology by compartmentalizing it into two distinct adapter modules: a domain adapter for domain-invariant feature extraction, and a task adapter with static parameters. The underlying architecture was based on AdapterFusion [121]. Another advancement, AdapterSoup [122], further optimized the adaptation process by utilizing a weight-averaging approach on domain adapters exclusively during the evaluation stage. Various techniques for domain adapter selection were investigated, including exhaustive combination, text clustering, and semantic similarity measures.

Adapters with underlying neural network architectures are commonly referred to as *neural adapters*. The seminal design of such adapters is attributed to Houlsby et al. [117] and consists of a sequential arrangement of down-projection, a GeLU non-linear activation function [123], and up-projection. These components are integrated with feed-forward layers to serve as the foundational architecture. Subsequent work by Bapna et al. [124] streamlined this structure, reducing it to a single hidden-layer feed-forward network while empirically demonstrating its efficacy in domain adaptation tasks. These adapter modules are strategically positioned after the multi-head attention and feed-forward layers within the transformer architecture. Variants of such neural adapters are colloquially termed as either bottleneck adapters or serial adapters; in the present paper, we employ the term "serial adapters" to refer specifically to the architecture described in [117].

Finally, Low-rank adaptation (LoRA) [125] is inspired by the observation that large language models reside on an intrinsic subspace [126], where model parameters are efficiently updated. Therefore, learning in this subspace significantly reduces the amount of parameters. LoRA modules implant learnable SVD blocks as the subspace

with a low matrix rank $r \ll d$, where d is the dimension of input data. The matrices are added in parallel to the pre-trained weights, thus keeping them frozen during the fine-tuning. Notably, LoRA shows superiority in further reducing the number of trained parameters and introducing no latency during inference.

5.2 Full-Parameter fine-tuning

As the name suggests, in the paradigm of full-parameter fine-tuning, all model parameters are subject to change during training. With a higher training cost than the PEFT, full-parameter fine-tuning can generally lead to better performance than the parameter-efficient methods [127]. However, this phenomenon may not hold true on a simple dataset (e.g., a dataset with a lack of language diversity) for a specific downstream task [128]. Since PEFT only trains a relatively small number of parameters, models trained via full-parameter fine-tuning methods have a greater learning capacity. Moreover, the convergence of PEFT is generally not as fast as that of full-parameter fine-tuning [129]. As for training cost, it is reported [127] that “using full-parameters fine-tuning requires about 3-5 times the time cost” of LoRA fine-tuning. GPU memory consumption is also a concern because updating all the parameters can be impractical when dealing with LLMs. The significantly higher training cost of full-parameter fine-tuning poses a challenge for researchers in choosing which method to use. As for memory cost during training, several optimization methods have been proposed, such as Gradient Checkpointing [130], Zero Redundancy Optimizer [15] and Flashattention [65].

To mitigate the cost of training, many recent works on full-parameter fine-tuning aim to optimize memory consumption [131, 132], which significantly reduces the barrier of this research. For example, a new optimizer called LOMO (**LOw-Memory Optimization**) was proposed [131] to combine gradient computation and parameter update in one training step in order to improve memory efficiency. Stochastic gradient descent (SGD) was adopted in this method, and a theoretical analysis was provided to show the effectiveness of SGD on fine-tuning all the parameters of LLMs. As a result, the full parameter fine-tuning of a 65B model requires less than 192GB GPU memory (“a single machine with $8 \times$ RTX 3090, each with 24GB memory”). LOMO presents a practical solution to train LLMs in resource-constrained scenarios. Another recently proposed optimizer called MeZO [132] estimates gradients using only two forward passes and fine-tunes LLMs “with the same memory footprint as inference”. Requiring 55GB GPU memory, it can train a 30B model via full-parameter fine-tuning. The HiFT method [133], a hierarchical fine-tuning strategy, addresses GPU memory constraints in full-parameter fine-tuning of language models by updating only a subset of model parameters at each step. This block-by-block update approach enables HiFT to achieve comparable performance to conventional full-parameter fine-tuning with significant memory savings. Notably, HiFT supports the fine-tuning of 7B models on devices with 24GB memory without additional memory-saving techniques. This method, compatible with various optimizers, shows potential as a scalable and efficient solution for large language model adaptation in memory-limited environments.

Researchers are also paying attention to data-centric knowledge injection [134] when adapting a general-purpose foundation model towards a specific domain such

as healthcare. The knowledge injection can be achieved by fine-tuning on domain-specific textbooks, publications, and instructions. As an identified drawback [135] of full-parameter fine-tuning, trained models can distort their pre-trained features and underperform on data distributions unseen during fine-tuning.

6 LLM inference

Inference in Large Language Models (LLMs) like the GPT series is a critical stage where trained models are applied to generate text, answer questions, or perform other language tasks based on their training. With the expansive size and complexity of these models, enhancing the efficiency of the inference process is essential. This section examines various techniques to optimize LLMs for inference, focusing on strategies that reduce computational load and memory usage while maintaining high-quality outputs. The approaches explored include model compression methods like pruning and quantization, and dynamic inference techniques that adaptively adjust computation based on input data. These methods are crucial for deploying LLMs in real-world applications, where resource constraints and performance requirements are key considerations.

6.1 Model compression

Model compression and acceleration are prevalent techniques in which a cumbersome, slow-performing model is optimized to produce a streamlined version. This refined model not only requires minimal storage—making it apt for mobile device deployment—but also operates with reduced latency. Moreover, initially training a sizable model and subsequently compressing it can enhance training efficiency and bolster its generalization capabilities.

6.1.1 Pruning

Sparsity, one of the longest-standing concepts in machine learning, was introduced to the neural network field as early as the 1980s [136]. It was picked up again for “modern” deep networks in the late 2010s, first under the name of Pruning, with the primary goal of reducing inference costs [137]. In general, pruning methods can be divided into two categories, i.e., *structured pruning* and *unstructured pruning*. Structured pruning targets higher-granularity structures, such as entire neurons, channels, layers, or rows/columns of weight matrices. Structured pruning results in a model with reduced size that retains its original architectural structure, making it more hardware-friendly for deployment. On the other hand, unstructured pruning involves removing individual weights or connections throughout the model based on certain criteria (e.g., smallest magnitude weights). Unstructured pruning produces a model with “holes” or sparse weight matrices, which require specialized software or hardware for efficient deployment. Recent research efforts have been devoted to combining LLMs with pruning techniques, aiming to tackle the substantial size and computational costs associated with LLMs.

Unstructured pruning. Unstructured pruning reduces the complexity of an LLM by eliminating particular parameters without taking into account its intrinsic organization. This method focuses on individual weights or neurons in the LLM, typically

by setting a threshold and nullifying parameters beneath it. Yet, by not respecting the overarching structure of the LLM, it leads to a model with a non-uniform sparse makeup. This non-uniformity necessitates unique compression methods to effectively store and compute the trimmed model. SparseGPT [138] represents a rapid unstructured pruning technique designed specifically for LLMs with hundreds of billions of parameters, allowing for operation within mere hours. Remarkably, it can reduce parameters by as much as 60% without compromising the model's performance significantly. To address the demanding weight update process of SparseGPT, Wanda [139] introduces a novel pruning criterion. Wanda assesses each weight by calculating the product of its magnitude and the norm of its related input activations, using an estimation from a concise calibration dataset. This criterion is used for intra-layer comparisons in linear layer outputs, facilitating the exclusion of less significant weights in LLMs. In [140], the author proposes a novel dynamic inference scheme, DynaTran, which prunes activations at runtime with low overhead, which improves the throughput of transformer inference. The author designed an application-specific integrated circuit (ASIC) based architecture called AccelTran to implement the DynaTran. Specifically, several hardware-aware designs are proposed including matrix tiling and various dataflow to improve data reuse. Bai et al proposed SparseLLM [141], a novel global pruning framework for LLMs. Unlike prior methods limited to layer-wise sparsity, SparseLLM achieves global pruning with a novel optimization design and demonstrates that global pruning can retain model accuracy while reducing resource demands, making it advantageous for deployment in large-scale, resource-constrained environments.

Structured pruning. Structured pruning involves the selective removal of an entire group of weights. The definition of ‘group’, which makes those amenable to hardware speedup, could refer to weight blocks, neurons, filters/channels, attention heads, or other dedicated fine-grained sparse patterns. [142] introduce LLM-Pruner, a pioneering framework tailored for structured pruning of LLMs offering task-agnostic compression and efficient data usage. LLM-Pruner integrates a dependency detection mechanism to identify interconnected structures in the model. It utilizes an effective importance estimation approach, combining both first-order data and estimated Hessian information. This approach streamlines the selection of prime groups for pruning, enhancing the compression procedure. [143] propose LoSparse (Low-Rank and Sparse approximation), a novel model compression technique that approximates a weight matrix by the sum of a low-rank matrix and a sparse matrix. Pruning enhances the diversity of low-rank approximations, and low-rank approximation prevents pruning from losing too many expressive neurons. [144] further considers pruning the hidden dimension (e.g., embedding layers, layer normalization) of LLM besides pruning the attention heads and feed-forward layers. [145] proposed a new structured compression approach for LLMs, called ZipLM, which provides state-of-the-art compression-vs-accuracy results, while guaranteeing to match a set of (achievable) target speedups on any given target hardware. Specifically, given a task, a model, an inference environment, as well as a set of speedup targets, ZipLM identifies and removes redundancies in the model through iterative structured shrinking of the model's weight matrices.

Contextual pruning While sparsity stands as a viable strategy to mitigate the burden in LLM inference, existing techniques either necessitate expensive retraining, compromise the LLM’s intrinsic learning capabilities, or fail to accelerate real-time performance on contemporary hardware. Zichang Liu et al. [146] postulate that the application of *contextual sparsity* — utilizing small, input-dependent sets of attention heads and MLP parameters to approximate the dense model’s output — can overcome these challenges. Their investigations confirm the presence of contextual sparsity and its potential for precise prediction, enabling us to leverage it to hasten LLM inference without sacrificing model quality or learning abilities in context. To capitalize on these findings, they introduce Deja Vu, a system proficient in dynamically predicting contextual sparsity using a cost-effective algorithm. This system, coupled with an asynchronous and hardware-optimized execution, significantly accelerates LLM inference times.

6.1.2 Quantization

The quantization-based approach aims to achieve substantial model compression at the cost of affordable loss of model accuracy. While the conventional method for representation learning adopts floating-point numbers, quantization converts them to fewer bits such as integers or other discrete numbers, making models more efficient regarding both memory and computation, especially suitable for deployment on resource-constrained devices. Although this might lead to the loss of model precision (or quantization error) to some extent, careful quantization techniques can achieve substantial model compression with only minimal accuracy degradation. Based on which module of the model the quantization is applied to, quantization-based methods can be classified into four scenarios: (1) weight quantization, (2) activation quantization, and (3) fixed-point quantization.

Weight quantization. The most popular practice for quantization is weight quantization, which compresses language models by representing model weights using fewer bits. For example, Lee et al. [147] jointly learned a common quantization grid size and the division factor for pre-trained weights and performed element-wise division on weights. Instead of quantizing all weights of the model which may lead to moderate-to-high quantization error, another natural thought is to identify and quantize weights that are not important. Some works perform weight quantization after the training process. For instance, Lin et al. [148] identified and preserved only 1% of salient weights by observing activation that can largely reduce the quantization error. Dettmers et al. [149] and Wei et al. [150] identified and isolated outlier weights that potentially lead to large quantization error via different techniques such as filtering sensitivity-based algorithm [149] or identifying asymmetric presentation and scaling down problematic channels [150]. Some work leverages either activation or model outliers sensitive to accuracy degrade from weight quantization. For instance, Kim et al. [151] also employed a sensitivity-based that searches for optimal bit precision assignment and stores outliers and sensitive weight values in an efficient sparse format. Lee et al. [152] studied how activation outliers can amplify the error in weight quantization and assign higher precision to the weights susceptible to quantization caused by activation outliers. Guo et al. [153] handled outlier values locally by sacrificing values next to outliers

(usually not important) to accommodate those important outliers. Liu et al. [154] focused on the weight quantization of generative models by applying distillation based on generations produced by the pre-trained model. Frantar et al. [155] proposed a one-shot weight quantization method based on approximated second-order information, reducing the bandwidth of the GPT model down to 3 or 4 bits per weight. Lin et al. [156] proposed DuQuant, a novel quantization strategy that employs rotation and permutation transformations to manage activation outliers more effectively, achieving state-of-the-art performance for low-bit weight-activation quantization on various large language models (LLMs). Shao et al. [157] proposed OmniQuant, a quantization method leveraging learnable weight clipping and equivalent transformations, effectively optimizing quantization for large language models while achieving state-of-the-art performance across various low-bit quantization settings. Some works achieve weight quantization during training. For example, Yang et al. [158] proposed dynamic stashing quantization that dynamically quantizes the intermediate results between forward and backward processes for a significant reduction of the memory traffic during training. Yang et al. [159] used low-rank tensor train and tensor-train matrix formats to represent the embedding tables and linear layers during training. Dettmers et al. [160] backpropagated gradients through a frozen, 4-bit quantized pre-trained language model into Low-Rank Adapters (LoRA) and performed double quantization by quantizing quantization constants. Wortsman et al. [161] accelerated and stabilized large language-vision models by reducing the weights to low-bit values, such as using 16-bit precision for weight gradient computation and int8 multiplications for the forward pass and layer input gradient computations. Other works approach to focus on the pre-trained model. Gong et al. [162] quantized the pre-trained model in a task-agnostic way to obtain a “pre-quantized” model before fine-tuning and froze most of the quantized weights in the “pre-quantized” model.

Activation quantization. In addition to weight quantization, other techniques such as activation quantization and fixed-point quantization have been employed to ease the heavy memory consumption handling LLMs. Activation quantization deals with quantizing the intermediate values (i.e., activations) that arise during model inference. For instance, Liu et al. [163] proposed a framework agnostic to the neural work architecture by approximating the gradient descent of activation compression training [164] via a linearized version. Liu et al. [154] not only performed weight quantization but also quantized activations to 6-bit precision.

Fixed-point quantization. Fixed-point quantization represents weights and activations using fixed-point arithmetic to reduce memory usage and accelerate computations. Yu et al. [165] pruned transformer-based language models to meet the GPU’s acceleration constraint of structured sparse patterns with FP16 type. Then the floating-point sparse model is quantized into a fixed-point one by quantization-aware training.

6.1.3 Knowledge distillation

The distillation of domain-specific knowledge from LLMs into more compact neural networks has emerged as a promising area. Considering the specialty of LLMs, recent techniques of Knowledge Distillation can be divided into two streams: 1) *White-box*

Knowledge Distillation: the teacher model’s parameters are available to use; 2) *Black-box Knowledge Distillation*: only the teacher model’s predictions are accessible.

White-box knowledge distillation. This approach not only aims to substantially decrease inference latency but also to amplify the effectiveness of specialized task-solving capabilities. A compelling example is the work by Muhamed et al., who ingeniously compressed a behemoth 1.5 billion-parameter white-box LLM into a far more manageable 70 million-parameter model. This was specifically engineered for optimizing Click-Through Rate (CTR) prediction tasks. They introduced an innovative architecture featuring twin-structured BERT-like encoders coupled with a fusion layer. This allowed for a seamless cross-architecture knowledge distillation from a single LLM, yielding superior performance metrics in both real-time online and controlled offline environments [166]. In a parallel vein, several studies [167–171] have also made strides in this field by incorporating a specialized knowledge distillation module during the fine-tuning process of LLMs. This results in a twofold benefit: accelerated convergence rates and more efficient utilization of computational resources. The distillation module intelligently leverages pre-trained model parameters to expedite the convergence process, while concurrently training a selective subset of parameters to effectively counteract the issues associated with model over-parameterization. Extending this concept further, additional works [172, 173] have ventured into the intricate process of distilling the nuanced chain-of-thought reasoning capabilities inherent in larger models into their smaller counterparts. This allows the miniaturized models to inherit a form of *cognitive reasoning* from their more oversized progenitors, thereby enhancing their overall utility and performance.

Black-box knowledge distillation. Another line of research on Knowledge Distillation focuses on the somewhat elusive task of distilling knowledge from “black-box” large language models (LLMs) like ChatGPT. In these cases, researchers are often limited to interacting only with the model’s predictions, without the luxury of directly accessing its internal parameters or architecture. This is a particularly challenging endeavor because the traditional methods of knowledge distillation, which often rely on structural similarities or parameter sharing between the teacher and student models, are rendered inapplicable. These types of works [174–178] have leveraged LLMs as a query generation machine that directly generate *high quality* instruction following queries (and answers) to fine-tune smaller LLMs (e.g., LLaMA). The obtained smaller LLMs exhibit a stronger instruction-following capability.

6.1.4 Low-rank approximation

Due to low memory cost, low-rank approximation has made the model compression more viable and practical. A common approach is singular value decomposition (SVD). For a low-rank matrix $A \in \mathbb{R}^{m \times n}$, where r is the rank of matrix A , there exists $U \in \mathbb{R}^{m \times r}$, $V \in \mathbb{R}^{n \times r}$ are two orthogonal matrices; $\sigma \in \mathbb{R}^{r \times r}$ is a diagonal matrix with only the non-zero singular values of A . Through SVD, we reduce the memory cost from $O(mn)$ to $O((m + n) \times r)$, which is a huge saving in many scenarios.

In general, any linear matrix can be approximated through SVD. [179] omit the diagonal matrix in SVD decomposition and encode the residue of the original matrix and approximated matrix to achieve better performance. [180] used low-rank matrices

to evaluate the parameter importance. They utilize low-rank matrices to formulate the optimization problem and solve it to get the approximation of the original parameter. [181] applied low-rank approximation to reduce quantization errors. They use low-rank decomposition to reduce error without a huge impact on the speed of inference of LLM. [182] achieved low-rank approximation through the observation that data of NLP task is always in low-rank subspace. They first decompose the matrix of Feed-forward propagation through SVD and solve the optimization problem to get the needed low-rank matrices. [183] and [184] use conduct decomposition for layers in the transformer and GPT-2 respectively through the Kronecker product. It is a new way of "multiplication" different from the traditional matrix multiplication. [185] utilize low-rank approximation to reduce the parameters of generative transformers up to 25%. The non-contextual embeddings will have far fewer features compared with contextual ones, which is a huge saving for large language models. [186] tackles the storage problem of large language models through low-rank approaches. They store embeddings in low-rank format to reduce the memory cost, making the deployment of LLM in edge devices possible. [187] introduces DLoRA, a distributed fine-tuning framework for large language models that enhances parameter efficiency and privacy. By offloading fine-tuning tasks between cloud and edge devices, DLoRA addresses the limitations of purely cloud or edge-based solutions, ensuring data privacy and reducing computation and communication costs. The Kill and Revive algorithm further optimizes performance by dynamically tuning only the most responsive parameters, achieving significant reductions in workload while maintaining accuracy on downstream tasks. [188] presents SplitLoRA, an efficient fine-tuning framework that combines split learning with federated learning to address large model training burdens. By partitioning the model, SplitLoRA reduces computational demands on client devices while maintaining model accuracy. This framework, which uses LoRA for parameter-efficient tuning, achieves faster convergence and lower resource use compared to traditional federated approaches, making it suitable for deployment in resource-limited environments. [189] introduces DEALRec, a data-efficient fine-tuning method for LLM-based recommendation systems. This approach optimizes few-shot fine-tuning by selecting influential samples that are representative of full data, balancing both influence and effort scores to maximize accuracy with minimal data. DEALRec, tested on three real-world datasets, achieved superior performance over full-data fine-tuning while significantly reducing computational costs, making it effective for dynamic recommendation environments. [190] explores post-training quantization (PTQ) as a solution for reducing memory and computational demands of large language models (LLMs). PTQ is applied across three types of tensors—Weights, Activations, and KV Cache—to optimize efficiency while assessing impact on model performance. The study evaluates models from 11 families, such as LLaMA2, Falcon, and Vicuna, across five task categories, including basic NLP, dialogue, and long-context tasks. Key findings suggest specific bit-width quantization strategies that balance performance and efficiency, revealing trends in performance degradation across tensor types and model sizes. This comprehensive evaluation serves as a guide for selecting quantization methods suited to different LLM applications and offers insights into optimizing model deployment under resource constraints. [191] introduces a new post-training quantization

(PTQ) method specifically designed for LLMs to run with integer-only operations, aiming to eliminate floating-point computations. Key components include Fully-Smooth Block-Reconstruction (FSBR) to stabilize inter-channel variations, Dynamic Integer-only MatMul (DI-MatMul) for dynamic quantization in matrix multiplication, and specialized integer-only non-linear operators like DI-ClippedSoftmax. This framework achieves significant inference efficiency while retaining accuracy comparable to floating-point models, demonstrating I-LLM’s potential for resource-limited deployments on edge devices. [192] introduces an innovative quantization framework designed to enable high-performance inference for LLMs under various bit-precision configurations. ABQ-LLM tackles challenges in quantized inference, such as performance degradation at low bit widths and limited support for non-standard precision formats on GPUs. Key innovations include distribution correction methods to handle quantization-induced distribution shifts and a bit balance strategy to reduce asymmetry issues in low-bit quantization (e.g., INT2). ABQ-LLM outperforms existing methods like SmoothQuant and I-LLM, demonstrating significant acceleration and memory efficiency, especially in configurations like W2A8 for LLaMA models.

6.2 Dynamic acceleration

In Section 6.1, we have introduced techniques for reducing the number of parameters in an LLM for inference acceleration. These methods are general and agnostic to input data, i.e., *static* for any given input sequence. However, there is another line of methods that aims to improve the efficiency of LLM inference without reducing the number of parameters. Such methods typically are specific to different input sequences and we term them as *dynamic acceleration* methods. In general, existing dynamic acceleration methods include 3 categories, i.e., *early exit*, *token pruning*, and *token parallelism*. Early exist accelerates model inference by terminating inference at a particular layer-based on some criteria, i.e., making an LLM shallower. On the other hand, token pruning accelerates inference by skipping some tokens for higher layers based on their importance, i.e., making an LLM input shorter. Last, token parallelism considers leveraging certain techniques or algorithms to generate multiple tokens in parallel (opposite to autoregressive fashion that generates each token sequentially).

6.2.1 Early exit

Early exit is an inference acceleration strategy used in neural networks by skipping the computation of certain layers. The rationale behind early exit is that simpler input samples usually require less calculation to make predictions [193–195]. Pioneering explorations on early exit often rely on defining their own early-exit criterion: DeeBERT [196] adapts entropy as its exit criterion; RightTool [197] adapts softmax scores of prediction as its exit criterion; PABEE [198] exit inference when the intermediate predictions of the internal classifiers remain unchanged consecutively. PCEE-BERT [199] proposes a hybrid early exit criterion that combines confident score with patience counter. In other words, PCEE-BERT will early exit when enough numbers of consecutive intermediate layers are confident. SkipBERT [200] accelerates inference by skipping the computation of shallow layers when precomputed text

chunks are met. The Higher layers can be further skipped using the early-exit criterion. Short-Cutting Transformer [201] suggests a linear transformation-based method to cast intermediate representations as final representations, thus bypassing the transformer computation in between. Short-Cutting Transformer adapts the same early exit strategy as in CALM [202], where the LLM early exits when the difference between the highest and the second highest probabilities is bigger than CALM’s confidence threshold. MuE [203] extends dynamic early exit strategy to multimodal LLMs. Unique challenges arise since existing early exit strategies can not directly apply to the widely-used unified multimodal architecture with both encoder and decoder, due to the difficulty of making exit decisions when dependencies between encoder and decoder exit. MuE proposes its exit criterion based on the layer-wise input similarity, inspired by the saturation observation [204].

6.2.2 Input pruning

Input Pruning explores the opportunity for the dynamic reduction of input sequence length to improve the Transformer’s computational efficiency. Its intuition is similar to the human being’s reading comprehension capability it does not read all words equally. Instead, some words are focused with more interest while others are skimmed. For Transformer models, this means adopting a dynamic computation budget for different input tokens according to their contents.

Existing input pruning works can be categorized into two classes based on token removal or retention criteria. The first class uses value-based scoring (e.g., attention) to identify unimportant tokens. For instance, SpAtten [205] ranks tokens using importance scores and retains the top-k highest-scoring tokens. LTP [206] improves PoWER-BERT by introducing a learnable layer-wise threshold, enabling adaptive pruning length. ToP [207] overcomes the limitation of inaccurate token importance ranking in the self-attention mechanism through a ranking-distilled token distillation technique, which distills effective token rankings from the final layer of unpruned models to early layers of pruned models.

The second class of token pruning methods inserts a prediction module before each transformer layer to provide a more accurate token importance score prediction. TR-BERT [208] introduces a dynamic mechanism for making decisions about skipping tokens. It is trained with reinforcement learning with a reward that promotes classifier confidence and penalizes the number of retained tokens. Transkimmer [209] is a notable example that inserts a 2-layer MLP network at each layer as the prediction module. However, the extra prediction module can also introduce considerable inference latency overhead, which is unfriendly on resource-limited devices. PuMer [210] proposed a token reduction framework that uses text-informed pruning and modality-aware merging strategies to progressively reduce the tokens of input image and text, improving model inference speed and reducing memory footprint. PuMer learns to keep salient image tokens related to the input text and merges similar textual and visual tokens by adding lightweight token reducer modules at several cross-modal layers in the Vision-Language model. Infor-Coef [211] proposes a model acceleration approach for large language models that incorporates dynamic token downsampling and static pruning, optimized by the information bottleneck loss. The token sampler,

which is similar to the MLP module of Transkimmer, is trained for downsampling the token length before the multi-head attention layer. SMART-TRIM [212] incorporates lightweight trimming modules (MLP layers) into the backbone to perform task-specific pruning on redundant inputs and parameters, without the need for additional pre-training or data augmentation. LLMLingua-2 [213] proposes a task-agnostic prompt compression method by formulating it as a token classification task. Despite its small size, it achieves notable speedups, reducing latency by 1.6x-2.9x with a compression ratio of 2x-5x, while preserving crucial information for effective prompt understanding across various downstream tasks. Compressed Context Memory (CCM) [214] introduces a dynamic context compression mechanism for language model inference by integrating lightweight LoRA during forward passes, achieving a memory-efficient solution for expanding context without fine-tuning the entire model. However, CCM-concat’s higher memory demands at later time steps may be challenging for memory-constrained environments. GRIFFIN [215], introduced in Prompt-prompted Adaptive Structured Pruning for Efficient LLM Generation, is a training-free and calibration-free pruning method targeting transformer feedforward blocks for faster, memory-efficient LLM inference. Exploiting the phenomenon of “flocking,” where neurons show similar activations across tokens in a sequence, GRIFFIN selects key neurons during the prompt phase, maintaining model performance even with 50% parameter reduction. Compared to magnitude pruning and MoEs, GRIFFIN achieves comparable efficiency without training overhead, showing 1.25x-1.29x speed-ups on Llama 2 and Gemma models across classification and generation tasks. LazyLLM [216] is a dynamic token pruning method aimed at improving LLM inference efficiency for long contexts. Unlike static pruning, LazyLLM selectively calculates key-value (KV) pairs only for tokens essential to predicting the next token at each generation step. By progressively pruning tokens during both the prefilling and decoding stages, LazyLLM reduces time-to-first-token (TTFT) and overall generation time without sacrificing model accuracy. Tests on the Llama 2 model show a $2.34 \times$ TTFT speedup on multi-document QA, validating LazyLLM’s capability to accelerate LLM inference efficiently without fine-tuning.

6.2.3 Token parallelism

Inference from large autoregressive models like Transformers is slow - decoding K tokens takes K serial runs of the model. Recent works proposed to leverage techniques such as *speculative execution* [217] to achieve parallel generation of multiple tokens instead of a sequential manner. Leviathan et al. [218] introduces “speculative decoding,” an algorithm that accelerates the sampling process from autoregressive models like Transformers by computing several tokens in parallel without altering the output distribution. This is achieved by utilizing approximation models (smaller than the original LLM) to generate speculative prefixes, which are then expanded by the larger target model, thereby accelerating the inference process without compromising the output quality. SpS [219] follows a similar idea and proposed speculative sampling, which generates multiple tokens per transformer call and uses a modified rejection sampling method. SpS maintains the output distribution while accelerating the process by 2 to 2.5 times without altering the model itself. Spector et al. [220] introduces an

enhanced algorithm called staged speculative decoding to expedite inference in large language models (LLMs), particularly in small-batch, on-device scenarios. The method consists of structuring the speculative batch as a tree to decrease generation costs and employing an additional stage of speculative decoding to boost performance. These improvements collectively yield a 3.16x reduction in single-batch decoding latency for a 762M parameter GPT-2-L model without compromising output quality.

7 System design

System design is critical in optimizing Large Language Models (LLMs) like the GPT series for efficient inference, particularly in resource-constrained environments. This section explores key strategies such as hardware offloading, which manages computational resources by leveraging different storage hierarchies, and collaborative inference, which pools resources for enhanced processing capabilities. It also examines the adaptation of LLMs for edge devices, highlighting the importance of system design in maximizing the efficiency and scalability of LLMs across various deployment scenarios.

7.1 Deployment optimization

Hardware offloading. Hardware offloading means transferring temporarily unneeded data in LLM from faster accelerators to slower yet ample primary and secondary storage, such as CPU memory and disk. These data are subsequently reloaded as needed. This method allows large LLMs to operate efficiently on GPUs with restricted memory capacity. However, the offloading and reloading processes inherently introduce significant communication overhead. Therefore, effective offloading strategy plays a crucial role in enhancing overall system efficiency. FlexGen [18] can achieve high throughput by developing a linear programming-based search algorithm that can identify the optimal offloading strategy within a defined search space of possible offloading strategies. FlexGen further improves throughput by compressing both the weights and KV cache for LLMs down to 4 bits. Additionally, FlexGen can be extended to a multi-GPU setting by adopting pipeline parallelism. DeepSpeed [17] introduced an innovative technique called ZeRO-Inference to minimize latency and enhance throughput. This approach involves pinning the model weights in either DRAM or NVMe and dynamically streaming each layer into GPU memory for computation as required. For a multi-GPUs environment, DeepSpeed leverages both tensor parallelism and pipeline parallelism to attain optimal performance. FastServe [221] designs the offloading strategy based on the priority of the jobs. Key-value tensors associated with lower-priority jobs are offloaded to host memory, while key-value tensors needed for imminent use are loaded in advance. LUT Tensor Core [222] introduces a novel software-hardware co-design to accelerate low-bit LLM inference through a LUT-based mixed-precision General Matrix Multiplication (mpGEMM) approach. This design leverages operator fusion to minimize precompute overhead, table symmetrization to reduce memory storage requirements, and a bit-serial architecture for flexible precision combinations, ultimately achieving superior compute density and energy efficiency compared to conventional Tensor Cores. BrainTransformers [223] implements Transformer-based LLMs using Spiking Neural Networks (SNN), introducing SNN-compatible components and

synaptic plasticity mechanisms, achieving competitive performance on benchmarks like MMLU (63.2) and GSM8K (76.3) while offering potential energy efficiency and biological plausibility. Ripple [224] optimizes LLM inference on smartphones by leveraging neuron co-activation to reorganize neuron placement in flash memory, achieving up to $5.93\times$ reductions in I/O latency through a two-stage solution combining offline co-activation pattern analysis and online caching strategies, addressing the intersection of sparsity-driven algorithms and storage-level system co-design. TorchTitan [225] introduces a PyTorch-native distributed training system with modular 3D parallelism, elastic scaling, and hardware-software co-optimization, achieving up to 30

Collaborative inference. Collaborative inference involves the cooperative effort of multiple users or systems working collectively to conduct inference tasks for LLMs. Each participant contributes their resources, such as computing power or data. This collaborative approach serves to mitigate the constraints of individual users or systems, ultimately leading to more efficient and accurate inference when dealing with LLMs. PETALS [226] is a system that facilitates collaborative inference and fine-tuning of LLMs through online collaboration among multiple users. Each participant in this system can take on the roles of a server, a client, or both. Servers host specific model layers and respond to client requests. Also, an 8-bit compression technology is used for lowering the computational resources requirement and improving efficiency.

7.2 Support infrastructure

Libraries. In this section, we introduce several famous frameworks for LLMs. Microsoft introduced DeepSpeed [17] as a cutting-edge deep learning optimization library that incorporates a range of innovative technologies, including ZeRO and 3D Parallelism. These technologies enable efficient and effective training, inference, and compression tasks. Megatron-LM [16], developed by NVIDIA, is another famous framework for LLMs. It implements a combination of model and data parallelism and focuses on multi-node pre-training of transformer-based models like GPT. DeepSpeed offers an enhanced version of Megatron-LM that includes additional features such as MoE model training and Curriculum Learning. Like DeepSpeed and Megatron-LM, Colossal-AI [227] provides multi-level parallelism strategies for large-scale distributed training. These strategies encompass Auto-Parallelism, data, tensor, pipeline, and sequence parallelism. Additionally, its heterogeneous memory management component enhances training efficiency in distributed environments with heterogeneous devices. Mesh-TensorFlow [84] is a user-friendly framework seamlessly integrated with TensorFlow, specializing in model parallelism with a primary emphasis on distributed tensor computations. GPT-NeoX [228] builds upon Megatron-LM and incorporates innovations from DeepSpeed. MaxText¹ is an open source open-source framework that is designed for Google Cloud TPUs. Alpa [88] can automatically parallelize users' single-device code on distributed clusters with data, operator, and pipeline parallelism. Transforming the Hybrid Cloud for Emerging AI Workloads [229] envisions a full-stack co-design of hybrid cloud systems incorporating generative and agentic AI, edge-to-cloud integration, quantum acceleration, and unified abstractions, aiming to create secure, efficient,

¹<https://github.com/google/maxtext>

and sustainable platforms that foster breakthroughs in AI-driven research and applications across academia, industry, and society. LLMServingSim [230] simulates LLM inference serving at the granularity of iterations to capture dynamic workload variations, leverages computation redundancies across decoder blocks to avoid repetitive simulations, and provides a flexible framework for exploring heterogeneous processor designs, achieving less than 14.7

Edge devices. In recent research, there is a growing interest in deploying LLMs on edge devices, which typically have limited computational resources. Xu et al. [231] explore the use of techniques like Low-Rank Adaptation and Noise Contrastive Estimation to reduce the memory demands of LLMs when running them on edge devices. Furthermore, they introduce a novel approach for reducing noise by minimizing data payload size within the setting of Differential Privacy and Federated Learning (FL). Woisetschläger et al. [232] gives a comprehensive analysis of the feasibility of conducting Federated Learning with LLMs on contemporary edge computing systems, offering a comparison to traditional data-centralized computing approaches. Shen et al. [233] leverages cloud-deployed LLMs to coordinate models that meet user needs and perform training via edge federated learning. EdgeFormer [234] interleave attention modules with a shared feed-forward network in the decoder layer to achieve cost-effective parameterization. ProFormer [235] utilizes the LSH projection layer to replace traditional embedding lookup tables, thereby mitigating the need for extensive memory resources. GhostBERT [236] employs 1-dimensional convolutions to generate additional features, thereby mitigating memory and computational expenses. SqueezeBERT [237] adopts optimization techniques used in computer vision networks, such as grouped convolutions, to achieve notable speedups. LiteTransformer [238] reduces the computation of the transformer base model by isolating local feature extraction from global feature extraction. MobileLLM [239] achieves significant accuracy improvements for sub-billion parameter models by prioritizing depth over width and leveraging techniques like weight-sharing and grouped-query attention to enhance performance with minimal memory overhead. EdgeShard [240] optimizes LLM inference by partitioning models across heterogeneous edge devices and cloud servers, leveraging collaborative edge computing to reduce latency and increase throughput. Any-Precision LLM [241] leverages post-training quantization and incremental upscaling to minimize deployment costs by supporting multiple quantized LLMs of varying bit-widths, all within a memory footprint comparable to a single model. The Breakthrough Memory Solutions [242] enhance LLM performance by integrating processing-in-memory (PIM) and processing-near-memory (PNM) technologies, significantly boosting memory bandwidth and reducing energy consumption for large-scale AI applications. MELTing point [243] enables efficient mobile execution of LLMs by leveraging a benchmarking infrastructure that traces memory and energy usage, optimizing for on-device inference across various mobile and edge devices. The proposed model in [244] reduces the communication overhead and memory consumption in federated learning by introducing partial embedding updates and low-rank adaptation, enabling efficient training of large-vocabulary models on resource-constrained devices. LLMS [245] reduces context switching latency in mobile LLM services by employing chunk-wise KV cache compression, recompute pipelining, and lifecycle management, enabling efficient large

language model execution on resource-constrained devices. LocMoE [246] optimizes large language model training by introducing locality-based routing and communication strategies, reducing training time and enhancing load balance without sacrificing model accuracy. JetMoE [247] reduces inference computation by leveraging sparse activation in both attention and feed-forward layers, activating only a subset of experts for each input token, leading to a 70% reduction in computation compared to dense models like Llama2-7B. Bai et al proposed FedSpaLLM [248], an innovative approach to pruning large language models in federated learning settings, addressing the unique challenges of model and system heterogeneity. Key contributions include an aggregation function based on the ℓ_0 -norm for managing diverse pruning masks across clients and a novel layer-sampling strategy that ensures efficient resource utilization while maintaining the accuracy of the global model. This framework allows for a flexible yet robust pruning process, enhancing model personalization and scalability in federated environments. FusionLLM [249] enables efficient decentralized training of large DNNs across geo-distributed GPUs by representing models as operator DAGs, using an OP-Fence scheduler, and implementing adaptive compression with AdaTopK, achieving up to 9.39x speedup in training ResNet-101 and GPT-2 on heterogeneous networked environments. Chat AI [250] seamlessly integrates HPC infrastructure and cloud-based web services for privacy-preserving, real-time LLM serving, leveraging Slurm’s batch scheduling to efficiently utilize GPU resources and incorporating an SSH ForceCommand-based circuit breaker for enhanced security.

7.3 Other systems

Tabi [251] proposes an inference system with a multi-level inference engine to reduce the inference latency of LLMs. Instead of using the same model for all the queries, Tabi avoids invoking costly LLMs by employing multiple DNNs to handle heterogeneous queries within a task. Z. Peng et al. [252] leverage min-hash technique to improve the efficiency and scalability of near-duplicate sequence search for LLMs.

8 Technique categorization by resources

In this section, we explore various techniques applied to Large Language Models (LLMs) to enhance their efficiency in using different resources. The focus is on five key resources: computation, memory, energy, financial cost, and network communication. Each technique discussed plays a vital role in optimizing resource efficiency for LLMs. The extent of their impact, whether direct or indirect, varies based on the resource in question. The table 2 provides a comprehensive mapping of these relationships.

8.1 Computation efficiency

Computation efficiency in LLMs is crucial for faster training and inference. Techniques like transformer architectures with approximated and hardware-aware attention directly enhance computation efficiency by reducing the complexity of operations. Approximated attention mechanisms, for instance, simplify the computationally

Main Category	Technique	Sub-Category	Computation	Memory	Energy	Money	Communication
Transformer Architecture §3.1	Approximated Attention		✓	✓	✓	✓	
	Hardware-aware Attention		✓	✓	✓	✓	
Non-transformer Architecture §3.2	Modular Network		✓	✓	✓	✓	
	Other Architecture		✓	✓	✓	✓	
Distributed Training §4.1.1	Data Parallelism		✓	✓	✓		
	Model Parallelism		✓	✓	✓		
Mixed-precision Training §4.1.2			✓	✓	✓		✓
Data Efficiency §4.2	Training Objective				✓	✓	
	Data Augmentation				✓	✓	
Parameter-efficient Fine-tuning §5.1	Adapter-based Fine-tuning		✓	✓			
	Masking-based Fine-tuning		✓	✓			
Full-parameter Fine-tuning §5.2			✓	✓	✓		
Pruning §6.1.1	Unstructured Pruning		✓	✓	✓		
	Structured Pruning		✓	✓	✓		
Quantization §6.1.2	Contextual Pruning		✓		✓		
	Weight Quantization		✓	✓	✓		✓
Activation Quantization			✓	✓	✓		
	Fixed-point Quantization		✓	✓	✓		
Knowledge Distillation §6.1.3	White-box distillation		✓	✓			
	Black-box distillation		✓	✓			
Low-rank Approximation §6.1.4			✓	✓			
Dynamic Inference §6.2	Early Exit		✓		✓	✓	
	Input Pruning		✓			✓	
Deployment Optimization §7.1	Token Parallelism		✓			✓	
	Hardware Offloading			✓		✓	
Collaborative Inference			✓	✓	✓	✓	

Table 2 Mapping of Resource Efficiency Techniques to Key Resources in Large Language Models: This table presents a detailed overview of various techniques employed in the optimization of LLMs, categorizing them by their main and sub-categories. Bold checkmarks (✓) indicate direct impacts and regular checkmarks (✓) indicate indirect impacts on the resources.

intensive attention calculations, thus speeding up the process. Hardware-aware optimizations tailor models to exploit specific hardware capabilities, leading to more efficient computation. Unstructured, structured, and contextual pruning also directly impacts computation efficiency by eliminating redundant computations through the removal of less important weights or neurons. Indirect impacts are observed in data parallelism and parameter-efficient fine-tuning, where the distributed workload and reduced parameter updates respectively contribute to overall computational efficiency, albeit as a secondary effect.

8.2 Memory efficiency

Memory efficiency is critical, particularly for deployment on resource-constrained devices. Techniques like pruning and quantization explicitly target memory efficiency by reducing the model size. Pruning methods eliminate unnecessary weights, and quantization reduces the precision of weights, both leading to substantial memory savings. Knowledge Distillation, where a smaller model is trained to mimic a larger one, also directly improves memory efficiency. Indirect contributions come from distributed training, where data and model parallelism effectively manage memory usage across multiple devices, reducing the burden on individual units.

8.3 Energy efficiency

Energy efficiency is increasingly important in the context of sustainable AI. Structured pruning and quantization are directly beneficial, as they reduce the number of operations and the data size, leading to lower energy consumption for both training and inference. Contextual pruning, which adapts the pruning process based on the context, also leads to energy savings by minimizing unnecessary computations. While primarily aimed at computational efficiency, techniques like approximated attention indirectly contribute to energy savings due to the reduced computational load.

8.4 Financial cost efficiency

Financial cost efficiency is an indirect but significant benefit of various resource-efficient techniques. Data efficiency methods, such as optimized training objectives and data augmentation, indirectly reduce costs by improving the effectiveness of the data used, leading to potentially shorter training times and less computational resource usage. Dynamic inference techniques like early exit and input pruning indirectly contribute to monetary efficiency by reducing the operational demands during the inference phase, which can lower the overall deployment costs. In addition to those techniques introduced earlier, some recent works exclusively minimize the monetary cost. For example, SpotServe [253] is a novel system for serving generative large language models (LLMs) on preemptible GPU instances in the cloud, which are more cost-effective but less stable than regular instances. SpotServe dynamically adjusts LLM parallelization to manage instance availability and workload fluctuations, optimizing for throughput, latency, and cost. It also employs a unique approach for instance migration, minimizing communication costs, and utilizes stateful inference

recovery to efficiently resume operations after preemptions, ultimately reducing costs by 54% and improving latency compared to existing systems.

8.5 Network communication efficiency

In distributed training environments, network communication efficiency becomes crucial. Mixed-precision training explicitly addresses this by reducing the size of data that needs to be communicated between processors, directly impacting the efficiency of data transfer. Techniques like weight quantization also have a direct impact by minimizing the data payload during communication. Collaborative inference, an indirect approach, enhances communication efficiency by distributing inference tasks in a manner that optimizes data transfer and processing across the network.

9 Benchmark and evaluation metrics

9.1 Evaluation metrics

Evaluating the resource efficiency of large language models (LLMs) involves considering a multifaceted range of metrics. We provide a comprehensive analysis of various metrics in this section. These metrics collectively offer a holistic understanding of the resource efficiency of large language models and are crucial for guiding model selection based on specific application requirements.

9.1.1 Computation

- **FLOPs** (Floating-point operations) represents the number of arithmetic operations on floating-point numbers, providing a quantifiable measure of the computation efficiency. In the context of LLMs, FLOPs capture the complexity of their internal computations, including attention mechanisms, feed-forward layers, and activation functions [254, 255].
- **Training time** refers to the total duration required to train an LLM, typically measured in wall-clock minutes, hours, or days [46, 65]. It reflects the model's complexity and reveals the efficiency of the training algorithms and hardware. Optimized algorithms and hardware can significantly reduce training time, making LLMs more accessible and affordable.
- **Inference time/latency** quantifies the time it takes for an LLM to generate an output after receiving an input. It is typically measured in wall-clock time or CPU/GPU/TPU clock time in milliseconds or seconds and can be accessed in two ways: 1) end-to-end latency [67] and 2) next token generation latency [256]. The inference time is crucial for evaluating the practical applicability of LLMs in real-world scenarios, particularly those with time-sensitive interactions.
- **Throughput** quantifies the model's efficiency in processing requests. It defines the rate at which LLMs can generate output tokens or complete tasks, typically measured in tokens per second (TPS) [67, 226] or queries per second [257, 258]. A high throughput indicates an LLM's ability to handle multiple requests quickly, making it suitable for high-volume applications like real-time chatbots or large-scale content generation.

- **Speed-up ratio** measures the improvement in inference speed compared to a baseline model, typically expressed as a multiple (e.g., 1.2x, 3.5x) [62, 63]. The ratio is calculated as the quotient of the baseline execution time to the time taken by the improved LLM or the quotient of the improved LLM’s throughput to that of the baseline model. The speed-up ratio draws a relative comparison between the two models, which offers a quantitative measure of progress in optimizing model architectures, training algorithms, and hardware optimizations.

9.1.2 Memory

- **Number of parameters** represents the number of adjustable variables in the LLM’s neural network. A higher number of model parameters generally indicates a more complex model with a greater capacity to learn and represent intricate patterns in the data [11, 74]. However, this complexity often comes at a cost of increased computational and memory requirements.
- **Model size** of an LLM is determined by the storage space required for storing the entire model. It directly correlates with the number of parameters and is also influenced by the model’s architecture, training data size, and any additional resources it needs to run. The model size is often measured by the peak memory usage during training or inference [49, 51], which can be easily obtained through system monitoring tools. Similar to the number of parameters, larger architectures, and more data have been able to continuously improve transformer models’ performances [42], but this can become a limiting factor, especially when dealing with extremely large models that may not fit into the memory of standard hardware.

9.1.3 Energy

- **Energy consumption** of an LLM is typically expressed in Watt-hours (Wh) or Joules (J), reflecting the electrical power used during the LLM’s lifecycle. [259] proposed a method to calculate the total power consumption by combining GPU, CPU, and DRAM consumption and multiplying it by the Power Usage Effectiveness (PUE), accounting for the additional energy required to support the entire compute infrastructure.
- **Carbon emission** captures the greenhouse gas emissions associated with the model’s energy usage. It can be calculated by multiplying the energy consumption (kWh) by the carbon intensity, the grams of carbon dioxide or equivalents emitted per kWh of energy used (gCO_{2eq}/kWh) by the local power grid.

CodeCarbon [260], carbontracker [261], and experiment-impact-tracker [262] are available software packages designed for real-time tracking of energy consumption and carbon emissions. Other tools like MLCO2 Impact [263] and LLMCarbon [264] leverage machine learning to predict the energy usage and carbon footprint before actual training, enabling more informed resource allocation.

However, focusing solely on training overlooks the full picture. It is encouraged to report the energy consumption and carbon footprint during all stages of an LLM’s lifecycle: while energy consumption during the training and development phases can be substantial, its significance may diminish when compared to the cumulative lifetime costs of inference if the model is utilized intensively in production. BLOOM [5]

highlights this holistic perspective, demonstrating the significant impact of life cycle assessment in accurately quantifying environmental costs.

9.1.4 Financial cost

While the financial cost of developing and deploying LLMs is rarely reported by researchers, understanding it is valuable for prioritizing research directions based on cost-effectiveness, encouraging collaboration to optimize resource allocation, and promoting responsible LLM development. We propose a novel metric *dollars per parameter* for this measure.

- **Dollars per parameter** is a scaled metric for the financial cost of an LLM. It is obtained by dividing the total cost of training (or running) the LLM by the number of parameters. It accounts for differences in model size and complexity, allowing for fairer comparisons of cost efficiency between different models. Researchers may consider reporting not only the cloud compute and electricity cost [259] and the hardware and software expenses associated with training and running, but also the personnel costs involved in data collection, model architecture, and fine-tuning.

9.1.5 Network communication

- **Communication volume** refers to the total amount of data (in megabytes, gigabytes, etc.) transmitted across the network during a specific LLM execution or training run [265]. Measuring it involves monitoring network traffic between connected nodes, often employing dedicated software or system logs. Monitoring the communication volume is important for distributed LLMs, as high communication costs can lead to bottlenecks, slowing down training and increasing infrastructure expenses.

9.1.6 Other metrics

- **Compression ratio** quantifies the reduction in size of the compressed model compared to the original model. It can be expressed in the percentage of size reduced [137, 144] or the percentage of weights remaining [143]. High compression ratios without significantly compromising the performance indicate better compression efficiency, which facilitates deployment across diverse platforms, including those with limited computational resources.
- **Loyalty** and **Fidelity** are two similar metrics proposed by [266] and [267] respectively to measure the resemblance between the teacher and student models in terms of both predictions consistency and predicted probability distributions alignment. For large teacher models, enhancements in fidelity lead to improvements in generalization: the student's ability to predict previously unseen, in-distribution data [267]. Achieving a high loyalty (or fidelity) is also important for preserving a model's fairness, as recent work showed that model compression can amplify existing algorithmic bias [268].
- **Robustness** is another important but rarely reported metric. [266] used two metrics: after-attack accuracy and query number to evaluate robustness. After-attack accuracy measures the model's post-attack performance, while the query number

reflects the complexity of the attack required to succeed. Evaluating the robustness of LLMs is crucial to ensure reliability in real-world applications, especially for compressed models, as [269] found smaller deep neural networks tend to be more vulnerable to adversarial attacks, where slight input modifications can manipulate their output.

- **Pareto optimality** [270] is achieved by striking an optimal balance between various competing factors. It offers a systematic approach to balance trade-offs between resource efficiency and task performance by identifying solutions that reside on the Pareto frontier [271] — the boundary that delineates optimal trade-offs. This concept is especially meaningful as it provides a principled framework for decision-making, allowing researchers to navigate the intricate landscape of LLMs and make informed choices. For example, [272] examined the trade-offs between cost and accuracy with the Pareto frontier, identifying the most practical real-world information retrieval models. [273] explored the Pareto frontier between performance and FLOPs to determine whether and how much a method achieves Pareto improvement.

9.2 Benchmarks

The evaluations of LLMs’ resource efficiency currently rely heavily on general NLP benchmarks², such as GLUE [275], SuperGLUE [276], WMT [277, 278], and SQuAD [279, 280]. While the existing general NLP benchmarks offer valuable insights into a model’s performance on various tasks, they often fail to capture the nuances of resource-efficient approaches. We present a few benchmarks that involve efficiency considerations; however, there is still a need for more comprehensive and specialized benchmarks for LLM efficiency evaluations. These benchmarks should go beyond traditional metrics and address the unique challenges associated with the efficient utilization of computational resources, memory, energy consumption, financial cost, and communication overhead.

- **Dynaboard** [258] is an open-source dynamic benchmark that allows users to submit NLP models to be evaluated in the cloud, enabling real-time interaction and a more comprehensive assessment of model quality. Submitted models are evaluated on a combination of datasets across four tasks: Natural Language Inference, Question Answering, Sentiment Analysis, and Hate Speech. In addition to accuracy, Dynaboard also collects additional metrics such as memory usage, throughput, fairness, and robustness. Models are ranked according to a novel utility-based aggregation of these metrics called Dynascore, which can be customized by leaderboard creators by assigning weights that reflect the relative importance of each metric.
- **EfficientQA** [281] is an open-domain Question Answering (QA) challenge at NeurIPS 2020³ that focuses on building accurate, memory-efficient QA systems. It promotes efficient memory usage through three restrained tracks based on model size and accuracy: the most accurate model under 6 GB, the most accurate model under 500 MB, and the smallest model that achieves 25% accuracy. The model size is measured as the Docker image size that contains the complete, self-contained question-answering system.

²For an extensive collection of general NLP benchmarks, refer to a comprehensive overview by [274].

³<https://efficientqa.github.io/>

- **SustaiNLP 2020**⁴ Shared Task [282] encourages participants to develop energy-efficient NLP models. It uses SuperGLUE [276] to access the model’s performance across eight diverse NLU tasks. In addition to standard SuperGLUE metrics, this challenge uniquely evaluates the energy consumption of each submission during inference using [262]’s experiment-impact-tracker.
- **ELUE** (Efficient Language Understanding Evaluation) [273] is a benchmark and platform designed to evaluate and compare the efficiency of various NLP models. It covers six NLP datasets spanning Sentiment Analysis, Natural Language Inference, Similarity, and Paraphrase tasks. ELUE supports online evaluation for model performance, average FLOPs, and the number of parameters. While Dynaboard [258] and EfficientQA [281] submissions require the containerized model along with its required environment, ELUE only requires the model definition file in Python, which is less costly for users to upload.
- **VLUE** (Vision-Language Understanding Evaluation) [283] is a multi-task multi-dimension benchmark for evaluating vision-language model (VLM)s. It covers a set of fundamental vision language tasks: Image-Text Retrieval, Visual Question Answering, Visual Reasoning, and Visual Grounding, and maintains an online platform and leaderboard for evaluation and comparison of vision-language pre-training (VLP) models. Notably, VLUE measures both performance and inference time and evaluates the efficiency-performance trade-off, providing a more comprehensive assessment of VLMs’ practical value.
- **Long-Range Arena** (LRA)⁵ [284] is a benchmark suite specifically designed for evaluating the performance of efficient Transformer models on long-context tasks. It features a variety of tasks that require reasoning over long contexts, ranging from 1,000 to 16,000 tokens in length. These tasks cover different modalities like text, natural language, synthetic images, and mathematical expressions, and involve various reasoning types like similarity, structural, and visual-spatial reasoning. Evaluations can be run under controlled resource constraints like memory budget or execution time limit. This forces models to be efficient within these limitations, highlighting their ability to optimize performance under real-world constraints.
- **Efficiency-aware MS MARCO** is a post-hoc leaderboard created by [272] for the MS MARCO information retrieval (IR) benchmark [285] to include efficiency metrics such as average per-query latency and the corresponding cost budget in addition to accuracy to provide a more holistic evaluation of IR systems. Following [258], [272] ranked the models using an aggregation of accuracy, latency, and cost, expressed as a Dynascore.

10 Open challenges and future directions

As the field of Large Language Models (LLMs) continues to advance, it faces a multitude of open challenges that pave the way for promising research directions. This section elaborates on these challenges and potential avenues for future exploration.

⁴<https://sites.google.com/view/sustainlp2020>

⁵<https://github.com/google-research/long-range-arena>

10.1 Managing resource type disagreements

A complex challenge in optimizing Large Language Models (LLMs) is the reconciliation of conflicting resource demands. Various optimization techniques introduce trade-offs between different performance metrics [146, 155, 173, 180]. For instance, some methods might enhance computational efficiency at the cost of increasing the number of model parameters or vice versa [125, 160]. Additionally, strategies that reduce computational operations could inadvertently lead to higher memory requirements due to the complexities involved in managing sparse data structures [138, 146]. The key challenge, therefore, is to develop a holistic optimization strategy for LLMs. This strategy should balance multiple objectives, including computational efficiency, parameter count, and memory usage, ensuring that improvements in one area do not disproportionately compromise performance in another. Achieving this balance is essential for the advancement of LLMs, making them both efficient and practical for diverse applications.

10.2 Combining techniques for resource efficiency

In the realm of Large Language Models (LLMs), a significant challenge lies in effectively amalgamating the diverse array of available methods to enhance overall resource efficiency. While numerous techniques exist to optimize different aspects of LLMs, there's a notable scarcity of research on how these methods can be cohesively combined. For instance, Ford et al. [286] propose a methodology to integrate energy efficiency analysis into software development cycles, which is relevant for the development of efficient LLMs. Similarly, Xie et al. [287] discuss the integration of resource allocation and task assignment for optimizing business processes, a concept that can be applied to the management of LLM resources. An integrative approach that systematically brings together various strategies could markedly improve the efficiency of these large models. This approach could draw from the insights of [288], who surveyed computational intelligence-based optimization and game theory approaches for resource allocation in computing environments. This could inform the development of strategies for efficient resource utilization in LLMs.

10.3 Standardized and unified evaluation

A critical challenge in the realm of LLMs is the absence of universally accepted benchmarks specifically tailored for evaluating the resource efficiency of these models. While several benchmarks exist for assessing aspects like model compression and acceleration [258, 281], they fall short of providing a comprehensive and consistent evaluation of resource efficiency. This shortcoming is due to varying factors such as differences in speed-up ratios, the number of parameters, accuracy levels, and even the type of hardware used in different studies. These discrepancies highlight the urgent need for standardized benchmarks that are designed with a focus on resource efficiency. Such benchmarks should facilitate direct comparisons and enable more accurate and holistic analyses of how various LLMs utilize computational resources, including aspects like

energy consumption, memory usage, and processing power. Establishing these benchmarks is essential for advancing the development of more resource-efficient LLMs, a key priority given the increasing size and complexity of these models.

10.4 Explainability and robustness

The pursuit of efficiency in Large Language Models (LLMs) brings to the fore concerns about their explainability and robustness, echoing issues identified in earlier research on pre-trained language models. For instance, while some techniques significantly improve performance in reasoning tasks, there is often a lack of clarity regarding the underlying reasons for their effectiveness [289, 290]. This obscurity highlights the need for integrating explainable approaches into the development of efficient LLMs. Such integration would address the dual needs of interpretability and simplified evaluation, thus enhancing the reliability and predictability of LLMs in real-world applications. The goal is to develop methods that not only optimize resource use but also maintain a level of transparency and resilience, ensuring that these models can be trusted and easily understood in various deployment scenarios.

10.5 AutoML for resource-efficient LLMs

The integration of Automated Machine Learning (AutoML) into the development of resource-efficient Large Language Models (LLMs) represents a burgeoning field of interest. Traditional methods for enhancing resource efficiency in LLMs, such as knowledge distillation, pruning, weight sharing, and low-rank factorization, typically rely on expert-driven heuristics and intricate manual interventions [146, 159, 180]. For instance, designing effective loss functions for knowledge distillation or determining saliency scores for pruning involves a considerable amount of human judgment and expertise [138, 173]. To mitigate this reliance on human input, there's a growing emphasis on applying techniques like Meta-Learning [291] and Neural Architecture Search (NAS) [292]. These AutoML strategies show promise in automating aspects of model optimization. By doing so, they could significantly reduce the need for manual hyperparameter tuning and bespoke model design, potentially leading to more efficient and easily customizable LLMs. This approach not only streamlines the optimization process but also opens avenues for more innovative and adaptable efficiency solutions in the realm of LLMs.

10.6 Edge computing with LLMs

Deploying Large Language Models (LLMs) in edge computing environments presents unique challenges due to the inherent limitations of edge devices. These devices often face constraints in terms of battery life, computational power, and memory resources [293, 294]. Additionally, issues such as data privacy and network latency further complicate their use [295]. To address these challenges, there is a need to develop LLM techniques that are not only resource-efficient but also mindful of privacy concerns. Key to this development is the ability to facilitate effective on-device training and operational capabilities of LLMs, making them viable for a range of practical applications in edge computing scenarios.

10.7 Theoretical insights into scaling laws

A crucial yet underexplored area in the realm of Large Language Models is the theoretical understanding of their scaling laws and generalization capabilities. A deeper insight into how LLMs' performance scales with their size and complexity is essential. Such understanding is pivotal in developing methods that are not just focused on model compression but are tailored to enhance the overall resource efficiency of LLMs. Gaining this knowledge is fundamental for researchers to more effectively explore and innovate within the design space of LLMs, ultimately leading to solutions that are both powerful and resource-efficient. This theoretical foundation is key to advancing the field and optimizing LLMs for a wider range of applications and environments.

11 Conclusion

In this survey, we have systematically explored the realm of resource-efficient Large Language Models (LLMs), offering a comprehensive view of the current state-of-the-art techniques and methodologies. We started by providing a foundational understanding of the challenges and necessities in developing resource-efficient LLMs in Section 2. This set the stage for a deeper investigation into various aspects of LLMs, including architecture design, pre-training, fine-tuning, inference, and system design.

In Section 3, we highlighted innovative approaches in LLM architecture that contribute to resource efficiency, emphasizing the significance of both transformer and non-transformer architectures. Sections 4, 5, and 6 delved into the nuances of pre-training, fine-tuning, and inference phases of LLMs, respectively, showcasing how each stage presents unique opportunities and challenges for resource optimization.

Our exploration in Section 7 underscored the importance of holistic system design in achieving resource efficiency, where both hardware and software aspects play a crucial role. In Section 8, we examined the practical applications and evaluations of these techniques, linking them back to the resource taxonomy established earlier in the survey.

The survey also shed light on the current benchmarks and evaluation metrics in Section 9, which are crucial for quantifying and comparing the efficiency of various LLMs. In Section 10, we identified key open challenges and future directions, pointing towards unexplored areas and potential breakthroughs that could further advance the field of resource-efficient LLMs.

As we conclude, it is evident that while significant progress has been made in developing resource-efficient LLMs, there remains a vast landscape of opportunities for further research and innovation. The insights and discussions presented in this survey aim to serve as a guiding framework for future work in this rapidly evolving field, with the ultimate goal of achieving highly efficient LLMs that are accessible and sustainable for a wide range of applications.

References

- [1] OpenAI, R.: Gpt-4 technical report. arxiv 2303.08774. View in Article 2, 13 (2023)

- [2] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., Lample, G.: LLaMA: Open and Efficient Foundation Language Models (2023)
- [3] Floridi, L., Chiriatti, M.: Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines* **30**, 681–694 (2020)
- [4] Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X.V., et al.: Opt: Open pre-trained transformer language models. arXiv preprint arXiv:2205.01068 (2022)
- [5] Scao, T.L., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., Castagné, R., Luccioni, A.S., Yvon, F., Gallé, M., et al.: Bloom: A 176b-parameter open-access multilingual language model. arXiv preprint arXiv:2211.05100 (2022)
- [6] Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H.W., Sutton, C., Gehrman, S., et al.: Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research* **24**(240), 1–113 (2023)
- [7] Chien, A.A., Lin, L., Nguyen, H., Rao, V., Sharma, T., Wijayawardana, R.: Reducing the carbon impact of generative ai inference (today and in 2035). In: Proceedings of the 2nd Workshop on Sustainable Computer Systems, pp. 1–7 (2023)
- [8] Samsi, S., Zhao, D., McDonald, J., Li, B., Michaleas, A., Jones, M., Bergeron, W., Kepner, J., Tiwari, D., Gadepally, V.: From words to watts: Benchmarking the energy costs of large language model inference. arXiv preprint arXiv:2310.03003 (2023)
- [9] Šakota, M., Peyrard, M., West, R.: Fly-swat or cannon? cost-effective language model choice via meta-modeling. arXiv preprint arXiv:2308.06077 (2023)
- [10] Kaddour, J., Harris, J., Mozes, M., Bradley, H., Raileanu, R., McHardy, R.: Challenges and Applications of Large Language Models (2023)
- [11] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. *Advances in neural information processing systems* **30** (2017)
- [12] Kaplan, J., McCandlish, S., Henighan, T., Brown, T.B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., Amodei, D.: Scaling laws for neural language models. arXiv preprint arXiv:2001.08361 (2020)
- [13] Ying, X.: An overview of overfitting and its solutions. In: *Journal of Physics: Conference Series*, vol. 1168, p. 022022 (2019). IOP Publishing

- [14] Tirumala, K., Markosyan, A., Zettlemoyer, L., Aghajanyan, A.: Memorization without overfitting: Analyzing the training dynamics of large language models. *Advances in Neural Information Processing Systems* **35**, 38274–38290 (2022)
- [15] Rajbhandari, S., Rasley, J., Ruwase, O., He, Y.: Zero: Memory optimizations toward training trillion parameter models. In: SC20: International Conference for High Performance Computing, Networking, Storage and Analysis, pp. 1–16 (2020). IEEE
- [16] Shoeybi, M., Patwary, M., Puri, R., LeGresley, P., Casper, J., Catanzaro, B.: Megatron-lm: Training multi-billion parameter language models using model parallelism. arXiv preprint arXiv:1909.08053 (2019)
- [17] Aminabadi, R.Y., Rajbhandari, S., Awan, A.A., Li, C., Li, D., Zheng, E., Ruwase, O., Smith, S., Zhang, M., Rasley, J., et al.: Deepspeed-inference: enabling efficient inference of transformer models at unprecedented scale. In: SC22: International Conference for High Performance Computing, Networking, Storage and Analysis, pp. 1–15 (2022). IEEE
- [18] Sheng, Y., Zheng, L., Yuan, B., Li, Z., Ryabinin, M., Chen, B., Liang, P., Ré, C., Stoica, I., Zhang, C.: Flexgen: High-throughput generative inference of large language models with a single gpu. In: International Conference on Machine Learning, pp. 31094–31116 (2023). PMLR
- [19] Deng, L., Li, G., Han, S., Shi, L., Xie, Y.: Model compression and hardware acceleration for neural networks: A comprehensive survey. *Proceedings of the IEEE* **108**(4), 485–532 (2020)
- [20] Rae, J.W., Borgeaud, S., Cai, T., Millican, K., Hoffmann, J., Song, F., Aslanides, J., Henderson, S., Ring, R., Young, S., et al.: Scaling language models: Methods, analysis & insights from training gopher. arXiv preprint arXiv:2112.11446 (2021)
- [21] 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)
- [22] Anil, R., Dai, A.M., Firat, O., Johnson, M., Lepikhin, D., Passos, A., Shakeri, S., Taropa, E., Bailey, P., Chen, Z., et al.: Palm 2 technical report. arXiv preprint arXiv:2305.10403 (2023)
- [23] Gunantara, N.: A review of multi-objective optimization: Methods and its applications. *Cogent Engineering* **5**(1), 1502242 (2018)
- [24] Menghani, G.: Efficient deep learning: A survey on making deep learning models smaller, faster, and better. *ACM Computing Surveys* (2021) <https://doi.org/10.1145/3578938>

- [25] Liu, Y., Han, T., Ma, S., Zhang, J., Yang, Y., Tian, J., He, H., Li, A., He, M., Liu, Z., et al.: Summary of chatgpt-related research and perspective towards the future of large language models. *Meta-Radiology*, 100017 (2023)
- [26] Yang, J., Jin, H., Tang, R., Han, X., Feng, Q., Jiang, H., Yin, B., Hu, X.: Harnessing the power of llms in practice: A survey on chatgpt and beyond. arXiv preprint arXiv:2304.13712 (2023)
- [27] Zhao, W.X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., et al.: A survey of large language models. arXiv preprint arXiv:2303.18223 (2023)
- [28] Zhang, C., Zhang, C., Zheng, S., Qiao, Y., Li, C., Zhang, M., Dam, S.K., Thwal, C.M., Tun, Y.L., Huy, L.L., et al.: A complete survey on generative ai (aigc): Is chatgpt from gpt-4 to gpt-5 all you need? arXiv preprint arXiv:2303.11717 (2023)
- [29] Cao, Y., Li, S., Liu, Y., Yan, Z., Dai, Y., Yu, P.S., Sun, L.: A comprehensive survey of ai-generated content (aigc): A history of generative ai from gan to chatgpt. arXiv preprint arXiv:2303.04226 (2023)
- [30] Ling, C., Zhao, X., Lu, J., Deng, C., Zheng, C., Wang, J., Chowdhury, T., Li, Y., Cui, H., Zhao, T., et al.: Domain specialization as the key to make large language models disruptive: A comprehensive survey. arXiv preprint arXiv:2305.18703 **2305** (2023)
- [31] Mialon, G., Dessì, R., Lomeli, M., Nalmpantis, C., Pasunuru, R., Raileanu, R., Rozière, B., Schick, T., Dwivedi-Yu, J., Celikyilmaz, A., et al.: Augmented language models: a survey. arXiv preprint arXiv:2302.07842 (2023)
- [32] Xu, C., McAuley, J.: A survey on model compression and acceleration for pre-trained language models. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, pp. 10566–10575 (2023)
- [33] Zhu, X., Li, J., Liu, Y., Ma, C., Wang, W.: A survey on model compression for large language models. arXiv preprint arXiv:2308.07633 (2023)
- [34] Chitty-Venkata, K.T., Mittal, S., Emani, M., Vishwanath, V., Soman, A.K.: A survey of techniques for optimizing transformer inference. *Journal of Systems Architecture*, 102990 (2023)
- [35] Tay, Y., Dehghani, M., Bahri, D., Metzler, D.: Efficient transformers: A survey. arXiv e-prints, 2009 (2020)
- [36] Fournier, Q., Caron, G.M., Aloise, D.: A practical survey on faster and lighter transformers. *ACM Computing Surveys* **55**(14s), 1–40 (2023)

- [37] Zhuang, B., Liu, J., Pan, Z., He, H., Weng, Y., Shen, C.: A survey on efficient training of transformers. arXiv preprint arXiv:2302.01107 (2023)
- [38] Cheng, Y., Wang, D., Zhou, P., Zhang, T.: A survey of model compression and acceleration for deep neural networks. arXiv preprint arXiv:1710.09282 (2017)
- [39] Long, X., Ben, Z., Liu, Y.: A survey of related research on compression and acceleration of deep neural networks. In: Journal of Physics: Conference Series, vol. 1213, p. 052003 (2019). IOP Publishing
- [40] Capra, M., Bussolino, B., Marchisio, A., Shafique, M., Masera, G., Martina, M.: An updated survey of efficient hardware architectures for accelerating deep convolutional neural networks. Future Internet **12**(7), 113 (2020)
- [41] Dhillesararao, P., Boppu, S., Manikandan, M.S., Cenkeramaddi, L.R.: Efficient hardware architectures for accelerating deep neural networks: Survey. IEEE Access (2022)
- [42] Kenton, J.D.M.-W.C., Toutanova, L.K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of naacL-HLT, vol. 1, p. 2 (2019)
- [43] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. The Journal of Machine Learning Research **21**(1), 5485–5551 (2020)
- [44] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020)
- [45] Kitaev, N., Kaiser, L., Levskaya, A.: Reformer: The efficient transformer. In: International Conference on Learning Representations (2020). <https://openreview.net/forum?id=rkgNKkHtvB>
- [46] Katharopoulos, A., Vyas, A., Pappas, N., Fleuret, F.: Transformers are rnns: Fast autoregressive transformers with linear attention. In: International Conference on Machine Learning, pp. 5156–5165 (2020). PMLR
- [47] Shen, Z., Zhang, M., Zhao, H., Yi, S., Li, H.: Efficient attention: Attention with linear complexities. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 3531–3539 (2021)
- [48] Zhai, S., Talbott, W., Srivastava, N., Huang, C., Goh, H., Zhang, R., Susskind, J.: An attention free transformer. arXiv preprint arXiv:2105.14103 (2021)

- [49] Rabe, M.N., Staats, C.: Self-attention does not need $o(n^2)$ memory. arXiv preprint arXiv:2112.05682 (2022)
- [50] Jang, H., Kim, J., Jo, J.-E., Lee, J., Kim, J.: Mnncfast: A fast and scalable system architecture for memory-augmented neural networks. In: Proceedings of the 46th International Symposium on Computer Architecture, pp. 250–263 (2019)
- [51] Zandieh, A., Han, I., Daliri, M., Karbasi, A.: Kdeformer: Accelerating transformers via kernel density estimation. arXiv preprint arXiv:2302.02451 (2023)
- [52] Ma, X., Zhou, C., Kong, X., He, J., Gui, L., Neubig, G., May, J., Zettlemoyer, L.: Mega: Moving average equipped gated attention. In: The Eleventh International Conference on Learning Representations (2023). <https://openreview.net/forum?id=qNLe3iq2El>
- [53] Wang, Y., Xiao, Z.: Loma: Lossless compressed memory attention. arXiv preprint arXiv:2401.09486 (2024)
- [54] He, Z., Feng, G., Luo, S., Yang, K., He, D., Xu, J., Zhang, Z., Yang, H., Wang, L.: Two stones hit one bird: Bilevel positional encoding for better length extrapolation. arXiv preprint arXiv:2401.16421 (2024)
- [55] Arora, S., Eyuboglu, S., Zhang, M., Timalsina, A., Alberti, S., Zinsley, D., Zou, J., Rudra, A., Ré, C.: Simple linear attention language models balance the recall-throughput tradeoff (2024). <https://arxiv.org/abs/2402.18668>
- [56] Huang, S., Song, Y., Zhou, J., Lin, Z.: Cluster-wise Graph Transformer with Dual-granularity Kernelized Attention (2024). <https://arxiv.org/abs/2410.06746>
- [57] Zhang, J., wei, J., Huang, H., Zhang, P., Zhu, J., Chen, J.: SageAttention: Accurate 8-Bit Attention for Plug-and-play Inference Acceleration (2024). <https://arxiv.org/abs/2410.02367>
- [58] Aguilera-Martos, I., Herrera-Poyatos, A., Luengo, J., Herrera, F.: Local Attention Mechanism: Boosting the Transformer Architecture for Long-Sequence Time Series Forecasting (2024). <https://arxiv.org/abs/2410.03805>
- [59] Mahmood, K., Huang, S.: Enhanced Computationally Efficient Long LoRA Inspired Perceiver Architectures for Auto-Regressive Language Modeling (2024). <https://arxiv.org/abs/2412.06106>
- [60] Yang, E.: Signformer is all you need: Towards Edge AI for Sign Language (2024). <https://arxiv.org/abs/2411.12901>
- [61] Peng, B., Alcaide, E., Anthony, Q., Albalak, A., Arcadinho, S., Cao, H., Cheng, X., Chung, M., Grella, M., GV, K.K., et al.: Rwkv: Reinventing rnns for the

- transformer era. arXiv preprint arXiv:2305.13048 (2023)
- [62] Wang, X., Xiong, Y., Wei, Y., Wang, M., Li, L.: Lightseq: A high performance inference library for transformers. NAACL-HLT 2021, 113 (2021)
- [63] NVIDIA: FasterTransformer: A Faster Transformer Framework. <https://github.com/NVIDIA/FasterTransformer> (2021)
- [64] Lefauveux, B., Massa, F., Liskovich, D., Xiong, W., Caggiano, V., Naren, S., Xu, M., Hu, J., Tintore, M., Zhang, S., Labatut, P., Haziza, D.: xFormers: A modular and hackable Transformer modelling library. <https://github.com/facebookresearch/xformers> (2022)
- [65] Dao, T., Fu, D., Ermon, S., Rudra, A., Ré, C.: Flashattention: Fast and memory-efficient exact attention with io-awareness. Advances in Neural Information Processing Systems **35**, 16344–16359 (2022)
- [66] Dao, T.: Flashattention-2: Faster attention with better parallelism and work partitioning. arXiv preprint arXiv:2307.08691 (2023)
- [67] Kwon, W., Li, Z., Zhuang, S., Sheng, Y., Zheng, L., Yu, C.H., Gonzalez, J.E., Zhang, H., Stoica, I.: Efficient memory management for large language model serving with pagedattention. In: Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles (2023)
- [68] Liu, Z., Zhao, C., Iandola, F., Lai, C., Tian, Y., Fedorov, I., Xiong, Y., Chang, E., Shi, Y., Krishnamoorthi, R., et al.: Mobilelm: Optimizing sub-billion parameter language models for on-device use cases. arXiv preprint arXiv:2402.14905 (2024)
- [69] Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G., Dean, J.: Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In: International Conference on Learning Representations (2016)
- [70] Lepikhin, D., Lee, H., Xu, Y., Chen, D., Firat, O., Huang, Y., Krikun, M., Shazeer, N., Chen, Z.: Gshard: Scaling giant models with conditional computation and automatic sharding. In: International Conference on Learning Representations (2020)
- [71] Fedus, W., Zoph, B., Shazeer, N.: Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. The Journal of Machine Learning Research **23**(1), 5232–5270 (2022)
- [72] Du, N., Huang, Y., Dai, A.M., Tong, S., Lepikhin, D., Xu, Y., Krikun, M., Zhou, Y., Yu, A.W., Firat, O., et al.: Glam: Efficient scaling of language models with mixture-of-experts. In: International Conference on Machine Learning, pp. 5547–5569 (2022). PMLR

- [73] Zhou, Y., Lei, T., Liu, H., Du, N., Huang, Y., Zhao, V., Dai, A.M., Le, Q.V., Laudon, J., *et al.*: Mixture-of-experts with expert choice routing. *Advances in Neural Information Processing Systems* **35**, 7103–7114 (2022)
- [74] Artetxe, M., Bhosale, S., Goyal, N., Mihaylov, T., Ott, M., Shleifer, S., Lin, X.V., Du, J., Iyer, S., Pasunuru, R., *et al.*: Efficient large scale language modeling with mixtures of experts. In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 11699–11732 (2022)
- [75] Clark, A., De Las Casas, D., Guy, A., Mensch, A., Paganini, M., Hoffmann, J., Damoc, B., Hechtman, B., Cai, T., Borgeaud, S., *et al.*: Unified scaling laws for routed language models. In: *International Conference on Machine Learning*, pp. 4057–4086 (2022). PMLR
- [76] Malach, E.: Auto-regressive next-token predictors are universal learners. arXiv preprint arXiv:2309.06979 (2023)
- [77] Poli, M., Massaroli, S., Nguyen, E., Fu, D.Y., Dao, T., Baccus, S., Bengio, Y., Ermon, S., Ré, C.: Hyena hierarchy: Towards larger convolutional language models. arXiv preprint arXiv:2302.10866 (2023)
- [78] Fu, D.Y., Arora, S., Grogan, J., Johnson, I., Eyuboglu, S., Thomas, A.W., Spector, B.F., Poli, M., Rudra, A., Re, C.: Monarch mixer: A simple sub-quadratic gemm-based architecture. In: *Thirty-seventh Conference on Neural Information Processing Systems* (2023)
- [79] Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752 (2023)
- [80] Sun, Y., Dong, L., Zhu, Y., Huang, S., Wang, W., Ma, S., Zhang, Q., Wang, J., Wei, F.: You only cache once: Decoder-decoder architectures for language models. arXiv preprint arXiv:2405.05254 (2024)
- [81] Zhu, R.-J., Zhang, Y., Siferman, E., Sheaves, T., Wang, Y., Richmond, D., Zhou, P., Eshraghian, J.K.: Scalable matmul-free language modeling. arXiv preprint arXiv:2406.02528 (2024)
- [82] Choe, W., Ji, Y., Lin, F.: RWKV-edge: Deeply Compressed RWKV for Resource-Constrained Devices (2024). <https://arxiv.org/abs/2412.10856>
- [83] Baines, M., Bhosale, S., Caggiano, V., Goyal, N., Goyal, S., Ott, M., Lefauveux, B., Liptchinsky, V., Rabbat, M., Sheffer, S., et al.: Fairscale: A general purpose modular pytorch library for high performance and large scale training (2021)
- [84] Shazeer, N., Cheng, Y., Parmar, N., Tran, D., Vaswani, A., Koanantakool, P., Hawkins, P., Lee, H., Hong, M., Young, C., et al.: Mesh-tensorflow: Deep learning for supercomputers. *Advances in neural information processing systems* **31**

(2018)

- [85] Huang, Y., Cheng, Y., Bapna, A., Firat, O., Chen, D., Chen, M., Lee, H., Ngiam, J., Le, Q.V., Wu, Y., et al.: Gpipe: Efficient training of giant neural networks using pipeline parallelism. *Advances in neural information processing systems* **32** (2019)
- [86] Narayanan, D., Harlap, A., Phanishayee, A., Seshadri, V., Devanur, N.R., Ganger, G.R., Gibbons, P.B., Zaharia, M.: Pipedream: Generalized pipeline parallelism for dnn training. In: *Proceedings of the 27th ACM Symposium on Operating Systems Principles*, pp. 1–15 (2019)
- [87] Kim, T., Kim, H., Yu, G.-I., Chun, B.-G.: Bpipe: memory-balanced pipeline parallelism for training large language models. In: *International Conference on Machine Learning*, pp. 16639–16653 (2023). PMLR
- [88] Zheng, L., Li, Z., Zhang, H., Zhuang, Y., Chen, Z., Huang, Y., Wang, Y., Xu, Y., Zhuo, D., Xing, E.P., et al.: Alpa: Automating inter-and {Intra-Operator} parallelism for distributed deep learning. In: *16th USENIX Symposium on Operating Systems Design and Implementation (OSDI 22)*, pp. 559–578 (2022)
- [89] Jiang, Z., Lin, H., Zhong, Y., Huang, Q., Chen, Y., Zhang, Z., Peng, Y., Li, X., Xie, C., Nong, S., et al.: {MegaScale}: Scaling large language model training to more than 10,000 {GPUs}. In: *21st USENIX Symposium on Networked Systems Design and Implementation (NSDI 24)*, pp. 745–760 (2024)
- [90] Yang, H., Zhou, J., Fu, Y., Wang, X., Roane, R., Guan, H., Liu, T.: Protrain: Efficient llm training via memory-aware techniques. arXiv preprint arXiv:2406.08334 (2024)
- [91] Micikevicius, P., Narang, S., Alben, J., Diamos, G., Elsen, E., Garcia, D., Ginsburg, B., Houston, M., Kuchaiev, O., Venkatesh, G., et al.: Mixed precision training. *The International Conference on Learning Representation* (2018)
- [92] Johnson, T.B., Guestrin, C.: Training deep models faster with robust, approximate importance sampling. *Advances in Neural Information Processing Systems* **31** (2018)
- [93] Katharopoulos, A., Fleuret, F.: Not all samples are created equal: Deep learning with importance sampling. In: *International Conference on Machine Learning*, pp. 2525–2534 (2018). PMLR
- [94] Paul, M., Ganguli, S., Dziugaite, G.K.: Deep learning on a data diet: Finding important examples early in training. *Advances in Neural Information Processing Systems* **34**, 20596–20607 (2021)
- [95] Sorscher, B., Geirhos, R., Shekhar, S., Ganguli, S., Morcos, A.: Beyond neural

scaling laws: beating power law scaling via data pruning. *Advances in Neural Information Processing Systems* **35**, 19523–19536 (2022)

- [96] Chen, D., Huang, Y., Ma, Z., Chen, H., Pan, X., Ge, C., Gao, D., Xie, Y., Liu, Z., Gao, J., Li, Y., Ding, B., Zhou, J.: Data-Juicer: A One-Stop Data Processing System for Large Language Models (2023)
- [97] Renduchintala, H.S.V.N.S.K., Killamsetty, K., Bhatia, S., Aggarwal, M., Ramakrishnan, G., Iyer, R., Krishnamurthy, B.: INGENIOUS: Using informative data subsets for efficient pre-training of language models. In: Bouamor, H., Pino, J., Bali, K. (eds.) *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 6690–6705. Association for Computational Linguistics, Singapore (2023). <https://doi.org/10.18653/v1/2023.findings-emnlp.445> . <https://aclanthology.org/2023.findings-emnlp.445>
- [98] Bukharin, A., Liu, T., Wang, S., Zuo, S., Gao, W., Yan, W., Zhao, T.: Machine learning force fields with data cost aware training. In: *Proceedings of the 40th International Conference on Machine Learning. ICML'23. JMLR.org*, ??? (2023)
- [99] Pan, R., Liu, X., Diao, S., Pi, R., Zhang, J., Han, C., Zhang, T.: Lisa: Layer-wise importance sampling for memory-efficient large language model fine-tuning. arXiv preprint arXiv:2403.17919 (2024)
- [100] Hao, X., Zhu, Y., Appalaraju, S., Zhang, A., Zhang, W., Li, B., Li, M.: Mixgen: A new multi-modal data augmentation. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*, pp. 379–389 (2023)
- [101] Hou, L., Cao, Q., Yuan, Y., Zhao, S., Ma, C., Pan, S., Wan, P., Wang, Z., Shen, H., Cheng, X.: Augmentation-aware self-supervision for data-efficient gan training. arXiv preprint arXiv:2205.15677 (2022)
- [102] Lu, J., Huang, W., Zheng, N., Zeng, X., Yeung, Y., Chen, X.: Improving end-to-end speech processing by efficient text data utilization with latent synthesis. In: Bouamor, H., Pino, J., Bali, K. (eds.) *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 4916–4928. Association for Computational Linguistics, Singapore (2023). <https://doi.org/10.18653/v1/2023.findings-emnlp.327> . <https://aclanthology.org/2023.findings-emnlp.327>
- [103] Wei, W., Ren, X., Tang, J., Wang, Q., Su, L., Cheng, S., Wang, J., Yin, D., Huang, C.: Llmrec: Large language models with graph augmentation for recommendation. In: *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pp. 806–815 (2024)
- [104] Ye, J., Xu, N., Wang, Y., Zhou, J., Zhang, Q., Gui, T., Huang, X.: Llm-da: Data augmentation via large language models for few-shot named entity recognition. arXiv preprint arXiv:2402.14568 (2024)

- [105] Fan, Z., He, S.: Efficient data learning for open information extraction with pre-trained language models. In: Bouamor, H., Pino, J., Bali, K. (eds.) Findings of the Association for Computational Linguistics: EMNLP 2023, pp. 13056–13063. Association for Computational Linguistics, Singapore (2023). <https://doi.org/10.18653/v1/2023.findings-emnlp.869> . <https://aclanthology.org/2023.findings-emnlp.869>
- [106] Song, K., Tan, X., Qin, T., Lu, J., Liu, T.-Y.: Mass: Masked sequence to sequence pre-training for language generation. In: International Conference on Machine Learning, pp. 5926–5936 (2019). PMLR
- [107] He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16000–16009 (2022)
- [108] Li, Y., Fan, H., Hu, R., Feichtenhofer, C., He, K.: Scaling language-image pre-training via masking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 23390–23400 (2023)
- [109] Jiang, H., He, P., Chen, W., Liu, X., Gao, J., Zhao, T.: SMART: Robust and efficient fine-tuning for pre-trained natural language models through principled regularized optimization. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 2177–2190 (2020)
- [110] Zaken, E.B., Goldberg, Y., Ravfogel, S.: Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, pp. 1–9 (2022)
- [111] Xu, R., Luo, F., Zhang, Z., Tan, C., Chang, B., Huang, S., Huang, F.: Raise a child in large language model: Towards effective and generalizable fine-tuning. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 9514–9528 (2021)
- [112] Zhang, H., Li, G., Li, J., Zhang, Z., Zhu, Y., Jin, Z.: Fine-tuning pre-trained language models effectively by optimizing subnetworks adaptively. Advances in Neural Information Processing Systems **35**, 21442–21454 (2022)
- [113] Yu, C., Jeoung, S., Kasi, A., Yu, P., Ji, H.: Unlearning bias in language models by partitioning gradients. In: Findings of the Association for Computational Linguistics: ACL 2023, pp. 6032–6048 (2023)
- [114] Abdurrahman, M.S., Elezabi, H., Xu, B.C.: Typhoon: Towards an effective task-specific masking strategy for pre-trained language models. arXiv preprint arXiv:2303.15619 (2023)

- [115] Hao, J., Sun, W., Xin, X., Meng, Q., Chen, Z., Ren, P., Ren, Z.: Meft: Memory-efficient fine-tuning through sparse adapter. arXiv preprint arXiv:2406.04984 (2024)
- [116] Valipour, M., Rezagholizadeh, M., Kobyzev, I., Ghodsi, A.: Dylora: Parameter efficient tuning of pre-trained models using dynamic search-free low-rank adaptation. arXiv preprint arXiv:2210.07558 (2022)
- [117] Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., Gelly, S.: Parameter-efficient transfer learning for nlp. In: International Conference on Machine Learning, pp. 2790–2799 (2019). PMLR
- [118] Hu, Z., Lan, Y., Wang, L., Xu, W., Lim, E.-P., Lee, R.K.-W., Bing, L., Poria, S.: Llm-adapters: An adapter family for parameter-efficient fine-tuning of large language models. arXiv preprint arXiv:2304.01933 (2023)
- [119] Zhang, R., Zheng, Y., Mao, X., Huang, M.: Unsupervised domain adaptation with adapter. arXiv preprint arXiv:2111.00667 (2021)
- [120] Malik, B., Kashyap, A.R., Kan, M.-Y., Poria, S.: Uadapter-efficient domain adaptation using adapters. In: Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pp. 2241–2255 (2023)
- [121] Pfeiffer, J., Kamath, A., Rücklé, A., Cho, K., Gurevych, I.: AdapterFusion: Non-destructive task composition for transfer learning
- [122] Chronopoulou, A., Peters, M.E., Fraser, A., Dodge, J.: Adaptersoup: Weight averaging to improve generalization of pretrained language models. In: Findings of the Association for Computational Linguistics: EACL 2023, pp. 2009–2018 (2023)
- [123] Hendrycks, D., Gimpel, K.: Gaussian error linear units (gelus). arXiv preprint arXiv:1606.08415 (2016)
- [124] Bapna, A., Firat, O.: Simple, scalable adaptation for neural machine translation. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 1538–1548 (2019)
- [125] Hu, E.J., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., *et al.*: Lora: Low-rank adaptation of large language models. In: International Conference on Learning Representations (2021)
- [126] Aghajanyan, A., Gupta, S., Zettlemoyer, L.: Intrinsic dimensionality explains the effectiveness of language model fine-tuning. In: Proceedings of the 59th Annual

Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp. 7319–7328 (2021)

- [127] Sun, X., Ji, Y., Ma, B., Li, X.: A comparative study between full-parameter and lora-based fine-tuning on chinese instruction data for instruction following large language model. arXiv preprint arXiv:2304.08109 (2023)
- [128] Razuvayevskaya, O., Wu, B., Leite, J.A., Heppell, F., Srba, I., Scarton, C., Bontcheva, K., Song, X.: Comparison between parameter-efficient techniques and full fine-tuning: A case study on multilingual news article classification. arXiv preprint arXiv:2308.07282 (2023)
- [129] Ding, N., Qin, Y., Yang, G., Wei, F., Yang, Z., Su, Y., Hu, S., Chen, Y., Chan, C.-M., Chen, W., *et al.*: Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature Machine Intelligence* **5**(3), 220–235 (2023)
- [130] Chen, T., Xu, B., Zhang, C., Guestrin, C.: Training deep nets with sublinear memory cost. arXiv preprint arXiv:1604.06174 (2016)
- [131] Lv, K., Yang, Y., Liu, T., Gao, Q., Guo, Q., Qiu, X.: Full Parameter Fine-tuning for Large Language Models with Limited Resources (2023)
- [132] Malladi, S., Gao, T., Nichani, E., Damian, A., Lee, J.D., Chen, D., Arora, S.: Fine-Tuning Language Models with Just Forward Passes (2023)
- [133] Liu, Y., Zhang, Y., Li, Q., Liu, T., Feng, S., Wang, D., Zhang, Y., Schütze, H.: Hift: A hierarchical full parameter fine-tuning strategy. arXiv preprint arXiv:2401.15207 (2024)
- [134] Wu, C., Lin, W., Zhang, X., Zhang, Y., Wang, Y., Xie, W.: PMC-LLaMA: Towards Building Open-source Language Models for Medicine (2023)
- [135] Kumar, A., Raghunathan, A., Jones, R., Ma, T., Liang, P.: Fine-tuning can distort pretrained features and underperform out-of-distribution. In: International Conference on Learning Representations (2022)
- [136] LeCun, Y., Denker, J., Solla, S.: Optimal brain damage. *Advances in neural information processing systems* **2** (1989)
- [137] Han, S., Mao, H., Dally, W.J.: Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. arXiv preprint arXiv:1510.00149 (2015)
- [138] Frantar, E., Alistarh, D.: Massive language models can be accurately pruned in one-shot. arXiv preprint arXiv:2301.00774 (2023)
- [139] Sun, M., Liu, Z., Bair, A., Kolter, J.Z.: A simple and effective pruning approach

- for large language models. arXiv preprint arXiv:2306.11695 (2023)
- [140] Tuli, S., Jha, N.K.: Acceltran: A sparsity-aware accelerator for dynamic inference with transformers. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (2023)
- [141] Bai, G., Li, Y., Ling, C., Kim, K., Zhao, L.: Gradient-free adaptive global pruning for pre-trained language models. arXiv preprint arXiv:2402.17946 (2024)
- [142] Ma, X., Fang, G., Wang, X.: Llm-pruner: On the structural pruning of large language models. arXiv preprint arXiv:2305.11627 (2023)
- [143] Li, Y., Yu, Y., Zhang, Q., Liang, C., He, P., Chen, W., Zhao, T.: Losparse: Structured compression of large language models based on low-rank and sparse approximation. arXiv preprint arXiv:2306.11222 (2023)
- [144] Tao, C., Hou, L., Bai, H., Wei, J., Jiang, X., Liu, Q., Luo, P., Wong, N.: Structured pruning for efficient generative pre-trained language models. In: Findings of the Association for Computational Linguistics: ACL 2023, pp. 10880–10895 (2023)
- [145] Kurtic, E., Frantar, E., Alistarh, D.: Ziplm: Hardware-aware structured pruning of language models. arXiv preprint arXiv:2302.04089 (2023)
- [146] Liu, Z., Wang, J., Dao, T., Zhou, T., Yuan, B., Song, Z., Shrivastava, A., Zhang, C., Tian, Y., Re, C., *et al.*: Deja vu: Contextual sparsity for efficient llms at inference time. In: International Conference on Machine Learning, pp. 22137–22176 (2023). PMLR
- [147] Lee, J.H., Kim, J., Kwon, S.J., Lee, D.: Flexround: Learnable rounding based on element-wise division for post-training quantization. Proceedings of the 40th International Conference on Machine Learning (2023)
- [148] Lin, J., Tang, J., Tang, H., Yang, S., Dang, X., Han, S.: Awq: Activation-aware weight quantization for llm compression and acceleration. arXiv preprint arXiv:2306.00978 (2023)
- [149] Dettmers, T., Svirschevski, R., Egiazarian, V., Kuznedelev, D., Frantar, E., Ashkboos, S., Borzunov, A., Hoefer, T., Alistarh, D.: Spqr: A sparse-quantized representation for near-lossless llm weight compression. arXiv preprint arXiv:2306.03078 (2023)
- [150] Wei, X., Zhang, Y., Li, Y., Zhang, X., Gong, R., Guo, J., Liu, X.: Outlier suppression+: Accurate quantization of large language models by equivalent and optimal shifting and scaling. arXiv preprint arXiv:2304.09145 (2023)
- [151] Kim, S., Hooper, C., Gholami, A., Dong, Z., Li, X., Shen, S., Mahoney,

- M.W., Keutzer, K.: Squeezellm: Dense-and-sparse quantization. arXiv preprint arXiv:2306.07629 (2023)
- [152] Lee, C., Jin, J., Kim, T., Kim, H., Park, E.: Owq: Lessons learned from activation outliers for weight quantization in large language models. arXiv preprint arXiv:2306.02272 (2023)
- [153] Guo, C., Tang, J., Hu, W., Leng, J., Zhang, C., Yang, F., Liu, Y., Guo, M., Zhu, Y.: Olive: Accelerating large language models via hardware-friendly outlier-victim pair quantization. In: Proceedings of the 50th Annual International Symposium on Computer Architecture, pp. 1–15 (2023)
- [154] Liu, Z., Oguz, B., Zhao, C., Chang, E., Stock, P., Mehdad, Y., Shi, Y., Krishnamoorthi, R., Chandra, V.: Llm-qat: Data-free quantization aware training for large language models. arXiv preprint arXiv:2305.17888 (2023)
- [155] Frantar, E., Ashkboos, S., Hoefer, T., Alistarh, D.: Gptq: Accurate post-training quantization for generative pre-trained transformers. The International Conference on Learning Representations (2023)
- [156] Lin, H., Xu, H., Wu, Y., Cui, J., Zhang, Y., Mou, L., Song, L., Sun, Z., Wei, Y.: Rotation and permutation for advanced outlier management and efficient quantization of llms. arXiv preprint arXiv:2406.01721 (2024)
- [157] Shao, W., Chen, M., Zhang, Z., Xu, P., Zhao, L., Li, Z., Zhang, K., Gao, P., Qiao, Y., Luo, P.: Omnidquant: Omnidirectionally calibrated quantization for large language models. arXiv preprint arXiv:2308.13137 (2023)
- [158] Yang, G., Lo, D., Mullins, R., Zhao, Y.: Dynamic stashing quantization for efficient transformer training. Findings of the Association for Computational Linguistics: EMNLP 2023 (2023)
- [159] Yang, Z., Choudhary, S., Kunzmann, S., Zhang, Z.: Quantization-aware and tensor-compressed training of transformers for natural language understanding. Interspeech (2023)
- [160] Dettmers, T., Pagnoni, A., Holtzman, A., Zettlemoyer, L.: Qlora: Efficient finetuning of quantized llms. arXiv preprint arXiv:2305.14314 (2023)
- [161] Wortsman, M., Dettmers, T., Zettlemoyer, L., Morcos, A., Farhadi, A., Schmidt, L.: Stable and low-precision training for large-scale vision-language models. arXiv preprint arXiv:2304.13013 (2023)
- [162] Gong, Z., Liu, J., Wang, Q., Yang, Y., Wang, J., Wu, W., Xian, Y., Zhao, D., Yan, R.: Prequant: A task-agnostic quantization approach for pre-trained language models. Findings of the Association for Computational Linguistics: ACL 2023 (2023)

- [163] Liu, X., Zheng, L., Wang, D., Cen, Y., Chen, W., Han, X., Chen, J., Liu, Z., Tang, J., Gonzalez, J., *et al.*: Gact: Activation compressed training for generic network architectures. In: International Conference on Machine Learning, pp. 14139–14152 (2022). PMLR
- [164] Evans, R.D., Aamodt, T.: Ac-gc: Lossy activation compression with guaranteed convergence. Advances in Neural Information Processing Systems **34**, 27434–27448 (2021)
- [165] Yu, C., Chen, T., Gan, Z.: Boost transformer-based language models with gpu-friendly sparsity and quantization. In: Findings of the Association for Computational Linguistics: ACL 2023, pp. 218–235 (2023)
- [166] Muhamed, A., Keivanloo, I., Perera, S., Mracek, J., Xu, Y., Cui, Q., Rajagopalan, S., Zeng, B., Chilimbi, T.: Ctr-bert: Cost-effective knowledge distillation for billion-parameter teacher models. In: NeurIPS Efficient Natural Language and Speech Processing Workshop (2021)
- [167] Vucetic, D., Tayaranian, M., Ziaeefard, M., Clark, J.J., Meyer, B.H., Gross, W.J.: Efficient fine-tuning of compressed language models with learners. arXiv preprint arXiv:2208.02070 (2022)
- [168] Marjieh, R., Sucholutsky, I., Rijn, P., Jacoby, N., Griffiths, T.L.: What language reveals about perception: Distilling psychophysical knowledge from large language models. arXiv preprint arXiv:2302.01308 (2023)
- [169] Azerbayev, Z., Ni, A., Schoelkopf, H., Radev, D.: Explicit knowledge transfer for weakly-supervised code generation. arXiv preprint arXiv:2211.16740 (2022)
- [170] Gu, Y., Dong, L., Wei, F., Huang, M.: Knowledge distillation of large language models. arXiv preprint arXiv:2306.08543 (2023)
- [171] Zhang, R., Shen, J., Liu, T., Liu, J., Bendersky, M., Najork, M., Zhang, C.: Do not blindly imitate the teacher: Using perturbed loss for knowledge distillation. arXiv preprint arXiv:2305.05010 (2023)
- [172] Shridhar, K., Stolfo, A., Sachan, M.: Distilling multi-step reasoning capabilities of large language models into smaller models via semantic decompositions. arXiv preprint arXiv:2212.00193 (2022)
- [173] Hsieh, C.-Y., Li, C.-L., Yeh, C.-K., Nakhost, H., Fujii, Y., Ratner, A., Krishna, R., Lee, C.-Y., Pfister, T.: Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. arXiv preprint arXiv:2305.02301 (2023)
- [174] Wu, M., Waheed, A., Zhang, C., Abdul-Mageed, M., Aji, A.F.: Lamini-lm: A diverse herd of distilled models from large-scale instructions. arXiv preprint

arXiv:2304.14402 (2023)

- [175] Peng, B., Li, C., He, P., Galley, M., Gao, J.: Instruction tuning with gpt-4. arXiv preprint arXiv:2304.03277 (2023)
- [176] Chiang, W.-L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J.E., et al.: Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. See <https://vicuna.lmsys.org> (accessed 14 April 2023) (2023)
- [177] Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., Liang, P., Hashimoto, T.B.: Stanford alpaca: An instruction-following llama model (2023)
- [178] Ling, C., Zhang, X., Zhao, X., Liu, Y., Cheng, W., Oishi, M., Osaki, T., Matsuda, K., Chen, H., Zhao, L.: Open-ended commonsense reasoning with unrestricted answer candidates. In: Findings of the Association for Computational Linguistics: EMNLP 2023, pp. 8035–8047 (2023)
- [179] Wang, H., Li, R., Jiang, H., Wang, Z., Tang, X., Bi, B., Cheng, M., Yin, B., Wang, Y., Zhao, T., et al.: Lighttoken: A task and model-agnostic lightweight token embedding framework for pre-trained language models. In: Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 2302–2313 (2023)
- [180] Zhang, M., Shen, C., Yang, Z., Ou, L., Yu, X., Zhuang, B., et al.: Pruning meets low-rank parameter-efficient fine-tuning. arXiv preprint arXiv:2305.18403 (2023)
- [181] Wu, X., Yao, Z., He, Y.: Zeroquant-fp: A leap forward in llms post-training w4a8 quantization using floating-point formats. arXiv preprint arXiv:2307.09782 (2023)
- [182] Chen, P., Yu, H.-F., Dhillon, I., Hsieh, C.-J.: Drone: Data-aware low-rank compression for large nlp models. Advances in neural information processing systems **34**, 29321–29334 (2021)
- [183] Tahaei, M.S., Charlaix, E., Nia, V.P., Ghodsi, A., Rezagholizadeh, M.: Kroneckerbert: Learning kronecker decomposition for pre-trained language models via knowledge distillation. arXiv preprint arXiv:2109.06243 (2021)
- [184] Edalati, A., Tahaei, M., Rashid, A., Nia, V.P., Clark, J.J., Rezagholizadeh, M.: Kronecker decomposition for gpt compression. arXiv preprint arXiv:2110.08152 (2021)
- [185] Reid, M., Marrese-Taylor, E., Matsuo, Y.: Subformer: Exploring weight sharing for parameter efficiency in generative transformers. arXiv preprint arXiv:2101.00234 (2021)

- [186] Xu, M., Xu, Y.L., Mandic, D.P.: Tensorgpt: Efficient compression of the embedding layer in llms based on the tensor-train decomposition. arXiv preprint arXiv:2307.00526 (2023)
- [187] Gao, C., Zhang, S.Q.: Dlora: Distributed parameter-efficient fine-tuning solution for large language model. arXiv preprint arXiv:2404.05182 (2024)
- [188] Lin, Z., Hu, X., Zhang, Y., Chen, Z., Fang, Z., Chen, X., Li, A., Vepakomma, P., Gao, Y.: Splitlora: A split parameter-efficient fine-tuning framework for large language models. arXiv preprint arXiv:2407.00952 (2024)
- [189] Lin, X., Wang, W., Li, Y., Yang, S., Feng, F., Wei, Y., Chua, T.-S.: Data-efficient fine-tuning for llm-based recommendation. In: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 365–374 (2024)
- [190] Li, S., Ning, X., Wang, L., Liu, T., Shi, X., Yan, S., Dai, G., Yang, H., Wang, Y.: Evaluating Quantized Large Language Models (2024). <https://arxiv.org/abs/2402.18158>
- [191] Hu, X., Cheng, Y., Yang, D., Yuan, Z., Yu, J., Xu, C., Zhou, S.: I-LLM: Efficient Integer-Only Inference for Fully-Quantized Low-Bit Large Language Models (2024). <https://arxiv.org/abs/2405.17849>
- [192] Zeng, C., Liu, S., Xie, Y., Liu, H., Wang, X., Wei, M., Yang, S., Chen, F., Mei, X.: ABQ-LLM: Arbitrary-Bit Quantized Inference Acceleration for Large Language Models (2024). <https://arxiv.org/abs/2408.08554>
- [193] Teerapittayanon, S., McDanel, B., Kung, H.-T.: Branchynet: Fast inference via early exiting from deep neural networks. In: 2016 23rd International Conference on Pattern Recognition (ICPR), pp. 2464–2469 (2016). IEEE
- [194] Jawahar, G., Sagot, B., Seddah, D.: What does bert learn about the structure of language? In: ACL 2019-57th Annual Meeting of the Association for Computational Linguistics (2019)
- [195] Rogers, A., Kovaleva, O., Rumshisky, A.: A primer in bertology: What we know about how bert works. Transactions of the Association for Computational Linguistics **8**, 842–866 (2021)
- [196] Xin, J., Tang, R., Lee, J., Yu, Y., Lin, J.: Deebert: Dynamic early exiting for accelerating bert inference. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 2246–2251 (2020)
- [197] Schwartz, R., Stanovsky, G., Swayamdipta, S., Dodge, J., Smith, N.A.: The right tool for the job: Matching model and instance complexities. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp.

- [198] Zhou, W., Xu, C., Ge, T., McAuley, J., Xu, K., Wei, F.: Bert loses patience: Fast and robust inference with early exit. *Advances in Neural Information Processing Systems* **33**, 18330–18341 (2020)
- [199] Zhang, Z., Zhu, W., Zhang, J., Wang, P., Jin, R., Chung, T.-S.: Pcee-bert: Accelerating bert inference via patient and confident early exiting. In: Findings of the Association for Computational Linguistics: NAACL 2022, pp. 327–338 (2022)
- [200] Wang, J., Chen, K., Chen, G., Shou, L., McAuley, J.: Skipbert: Efficient inference with shallow layer skipping. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 7287–7301 (2022)
- [201] Yom Din, A., Karidi, T., Choshen, L., Geva, M.: Jump to conclusions: Short-cutting transformers with linear transformations. arXiv preprint arXiv:2303.09435 (2023)
- [202] Schuster, T., Fisch, A., Gupta, J., Dehghani, M., Bahri, D., Tran, V., Tay, Y., Metzler, D.: Confident adaptive language modeling. *Advances in Neural Information Processing Systems* **35**, 17456–17472 (2022)
- [203] Tang, S., Wang, Y., Kong, Z., Zhang, T., Li, Y., Ding, C., Wang, Y., Liang, Y., Xu, D.: You need multiple exiting: Dynamic early exiting for accelerating unified vision language model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10781–10791 (2023)
- [204] Geva, M., Caciularu, A., Wang, K., Goldberg, Y.: Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pp. 30–45 (2022)
- [205] Wang, H., Zhang, Z., Han, S.: Spatten: Efficient sparse attention architecture with cascade token and head pruning. In: 2021 IEEE International Symposium on High-Performance Computer Architecture (HPCA), pp. 97–110 (2021). IEEE
- [206] Kim, S., Shen, S., Thorsley, D., Gholami, A., Kwon, W., Hassoun, J., Keutzer, K.: Learned token pruning for transformers. In: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 784–794 (2022)
- [207] Li, J., Zhang, L.L., Xu, J., Wang, Y., Yan, S., Xia, Y., Yang, Y., Cao, T., Sun, H., Deng, W., et al.: Constraint-aware and ranking-distilled token pruning for efficient transformer inference. arXiv preprint arXiv:2306.14393 (2023)

- [208] Ye, D., Lin, Y., Huang, Y., Sun, M.: Tr-bert: Dynamic token reduction for accelerating bert inference. arXiv preprint arXiv:2105.11618 (2021)
- [209] Guan, Y., Li, Z., Leng, J., Lin, Z., Guo, M.: Transkimmer: Transformer learns to layer-wise skim. arXiv preprint arXiv:2205.07324 (2022)
- [210] Cao, Q., Paranjape, B., Hajishirzi, H.: Pumer: Pruning and merging tokens for efficient vision language models. arXiv preprint arXiv:2305.17530 (2023)
- [211] Tan, W.: Infor-coef: Information bottleneck-based dynamic token downsampling for compact and efficient language model. arXiv preprint arXiv:2305.12458 (2023)
- [212] Wang, Z., Chen, J., Zhou, W., Liu, M., Qin, B.: Smarttrim: Adaptive tokens and parameters pruning for efficient vision-language models. arXiv preprint arXiv:2305.15033 (2023)
- [213] Pan, Z., Wu, Q., Jiang, H., Xia, M., Luo, X., Zhang, J., Lin, Q., Rühle, V., Yang, Y., Lin, C.-Y., et al.: Llmlingua-2: Data distillation for efficient and faithful task-agnostic prompt compression. arXiv preprint arXiv:2403.12968 (2024)
- [214] Kim, J.-H., Yeom, J., Yun, S., Song, H.O.: Compressed context memory for online language model interaction. arXiv preprint arXiv:2312.03414 (2023)
- [215] Dong, H., Chen, B., Chi, Y.: Prompt-prompted Adaptive Structured Pruning for Efficient LLM Generation (2024). <https://arxiv.org/abs/2404.01365>
- [216] Fu, Q., Cho, M., Merth, T., Mehta, S., Rastegari, M., Najibi, M.: LazyLLM: Dynamic Token Pruning for Efficient Long Context LLM Inference (2024). <https://arxiv.org/abs/2407.14057>
- [217] Burton, F.W.: Speculative computation, parallelism, and functional programming. IEEE Transactions on Computers **100**(12), 1190–1193 (1985)
- [218] Leviathan, Y., Kalman, M., Matias, Y.: Fast inference from transformers via speculative decoding. In: International Conference on Machine Learning, pp. 19274–19286 (2023). PMLR
- [219] Chen, C., Borgeaud, S., Irving, G., Lespiau, J.-B., Sifre, L., Jumper, J.: Accelerating large language model decoding with speculative sampling. arXiv preprint arXiv:2302.01318 (2023)
- [220] Spector, B., Re, C.: Accelerating llm inference with staged speculative decoding. arXiv preprint arXiv:2308.04623 (2023)
- [221] Wu, B., Zhong, Y., Zhang, Z., Huang, G., Liu, X., Jin, X.: Fast distributed inference serving for large language models. arXiv preprint arXiv:2305.05920 (2023)

- [222] Mo, Z., Wang, L., Wei, J., Zeng, Z., Cao, S., Ma, L., Jing, N., Cao, T., Xue, J., Yang, F., et al.: Lut tensor core: Lookup table enables efficient low-bit llm inference acceleration. arXiv preprint arXiv:2408.06003 (2024)
- [223] Tang, Z., Zhu, E.: BrainTransformers: SNN-LLM (2024). <https://arxiv.org/abs/2410.14687>
- [224] Wang, T., Fan, R., Huang, M., Hao, Z., Li, K., Cao, T., Lu, Y., Zhang, Y., Ren, J.: Ripple: Accelerating LLM Inference on Smartphones with Correlation-Aware Neuron Management (2024). <https://arxiv.org/abs/2410.19274>
- [225] Liang, W., Liu, T., Wright, L., Constable, W., Gu, A., Huang, C.-C., Zhang, I., Feng, W., Huang, H., Wang, J., Purandare, S., Nadathur, G., Idreos, S.: TorchTitan: One-stop PyTorch native solution for production ready LLM pre-training (2024). <https://arxiv.org/abs/2410.06511>
- [226] Borzunov, A., Baranchuk, D., Dettmers, T., Ryabinin, M., Belkada, Y., Chumachenko, A., Samygin, P., Raffel, C.: Petals: Collaborative inference and fine-tuning of large models. arXiv preprint arXiv:2209.01188 (2022)
- [227] Li, S., Liu, H., Bian, Z., Fang, J., Huang, H., Liu, Y., Wang, B., You, Y.: Colossal-ai: A unified deep learning system for large-scale parallel training. In: Proceedings of the 52nd International Conference on Parallel Processing, pp. 766–775 (2023)
- [228] Andonian, A., Anthony, Q., Biderman, S., Black, S., Gali, P., Gao, L., Hallahan, E., Levy-Kramer, J., Leahy, C., Nestler, L., Parker, K., Pieler, M., Phang, J., Purohit, S., Schoelkopf, H., Stander, D., Songz, T., Tigges, C., Thérien, B., Wang, P., Weinbach, S.: GPT-NeoX: Large Scale Autoregressive Language Modeling in PyTorch (2023). <https://doi.org/10.5281/zenodo.5879544> . <https://www.github.com/eleutherai/gpt-neox>
- [229] Chen, D., Youssef, A., Pendse, R., Schleife, A., Clark, B.K., Hamann, H., He, J., Laino, T., Varshney, L., Wang, Y., Sil, A., Jabbarvand, R., Xu, T., Kindratenko, V., Costa, C., Adve, S., Mendis, C., Zhang, M., Núñez-Corrales, S., Ganti, R., Srivatsa, M., Kim, N.S., Torrellas, J., Huang, J., Seelam, S., Nahrstedt, K., Abdelzaher, T., Eilam, T., Zhao, H., Manica, M., Iyer, R., Hirzel, M., Adve, V., Marinov, D., Franke, H., Tong, H., Ainsworth, E., Zhao, H., Vasisht, D., Do, M., Oliveira, F., Pacifici, G., Puri, R., Nagpurkar, P.: Transforming the Hybrid Cloud for Emerging AI Workloads (2024). <https://arxiv.org/abs/2411.13239>
- [230] Cho, J., Kim, M., Choi, H., Heo, G., Park, J.: Llmservingsim: A hw/sw co-simulation infrastructure for llm inference serving at scale. In: 2024 IEEE International Symposium on Workload Characterization (IISWC), pp. 15–29. IEEE, ??? (2024). <https://doi.org/10.1109/iiswc63097.2024.00012> . <http://dx.doi.org/10.1109/IISWC63097.2024.00012>

- [231] Xu, M., Song, C., Tian, Y., Agrawal, N., Granqvist, F., Dalen, R., Zhang, X., Argueta, A., Han, S., Deng, Y., *et al.*: Training large-vocabulary neural language models by private federated learning for resource-constrained devices. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1–5 (2023). IEEE
- [232] Woisetschläger, H., Isenko, A., Wang, S., Mayer, R., Jacobsen, H.-A.: Federated fine-tuning of llms on the very edge: The good, the bad, the ugly. arXiv preprint arXiv:2310.03150 (2023)
- [233] Shen, Y., Shao, J., Zhang, X., Lin, Z., Pan, H., Li, D., Zhang, J., Letaief, K.B.: Large language models empowered autonomous edge ai for connected intelligence. arXiv preprint arXiv:2307.02779 (2023)
- [234] Ge, T., Chen, S.-Q., Wei, F.: Edgeformer: A parameter-efficient transformer for on-device seq2seq generation. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pp. 10786–10798 (2022)
- [235] Sankar, C., Ravi, S., Kozareva, Z.: Proformer: Towards on-device lsh projection based transformers. In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pp. 2823–2828 (2021)
- [236] Huang, Z., Hou, L., Shang, L., Jiang, X., Chen, X., Liu, Q.: Ghostbert: Generate more features with cheap operations for bert. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp. 6512–6523 (2021)
- [237] Iandola, F., Shaw, A., Krishna, R., Keutzer, K.: Squeezebert: What can computer vision teach nlp about efficient neural networks? In: Proceedings of SustaiNLP: Workshop on Simple and Efficient Natural Language Processing, pp. 124–135 (2020)
- [238] Wu, Z., Liu, Z., Lin, J., Lin, Y., Han, S.: Lite transformer with long-short range attention. arXiv preprint arXiv:2004.11886 (2020)
- [239] Liu, Z., Zhao, C., Iandola, F., Lai, C., Tian, Y., Fedorov, I., Xiong, Y., Chang, E., Shi, Y., Krishnamoorthi, R., Lai, L., Chandra, V.: MobileLLM: Optimizing Sub-billion Parameter Language Models for On-Device Use Cases (2024). <https://arxiv.org/abs/2402.14905>
- [240] Zhang, M., Cao, J., Shen, X., Cui, Z.: EdgeShard: Efficient LLM Inference via Collaborative Edge Computing (2024). <https://arxiv.org/abs/2405.14371>
- [241] Park, Y., Hyun, J., Cho, S., Sim, B., Lee, J.W.: Any-Precision LLM: Low-Cost Deployment of Multiple, Different-Sized LLMs (2024). <https://arxiv.org/abs/>

- [242] Kim, B., Cha, S., Park, S., Lee, J., Lee, S., Kang, S.-h., So, J., Kim, K., Jung, J., Lee, J.-G., Lee, S., Paik, Y., Kim, H., Kim, J.-S., Lee, W.-J., Ro, Y., Cho, Y., Kim, J.H., Song, J., Yu, J., Lee, S., Cho, J., Sohn, K.: The breakthrough memory solutions for improved performance on llm inference. *IEEE Micro* **44**(3), 40–48 (2024) <https://doi.org/10.1109/MM.2024.3375352>
- [243] Laskaridis, S., Katevas, K., Minto, L., Haddadi, H.: MELTing point: Mobile Evaluation of Language Transformers (2024). <https://arxiv.org/abs/2403.12844>
- [244] Xu, M., Song, C., Tian, Y., Agrawal, N., Granqvist, F., Dalen, R., Zhang, X., Argueta, A., Han, S., Deng, Y., Liu, L., Walia, A., Jin, A.: Training Large-Vocabulary Neural Language Models by Private Federated Learning for Resource-Constrained Devices (2022). <https://arxiv.org/abs/2207.08988>
- [245] Yin, W., Xu, M., Li, Y., Liu, X.: LLM as a System Service on Mobile Devices (2024). <https://arxiv.org/abs/2403.11805>
- [246] Li, J., Sun, Z., He, X., Zeng, L., Lin, Y., Li, E., Zheng, B., Zhao, R., Chen, X.: LocMoE: A Low-Overhead MoE for Large Language Model Training (2024). <https://arxiv.org/abs/2401.13920>
- [247] Shen, Y., Guo, Z., Cai, T., Qin, Z.: JetMoE: Reaching Llama2 Performance with 0.1M Dollars (2024). <https://arxiv.org/abs/2404.07413>
- [248] Bai, G., Li, Y., Li, Z., Zhao, L., Kim, K.: Fedspallm: Federated pruning of large language models. arXiv preprint arXiv:2410.14852 (2024)
- [249] Tang, Z., Kang, X., Yin, Y., Pan, X., Wang, Y., He, X., Wang, Q., Zeng, R., Zhao, K., Shi, S., Zhou, A.C., Li, B., He, B., Chu, X.: FusionLLM: A Decentralized LLM Training System on Geo-distributed GPUs with Adaptive Compression (2024). <https://arxiv.org/abs/2410.12707>
- [250] Doosthosseini, A., Decker, J., Nolte, H., Kunkel, J.M.: Chat AI: A Seamless Slurm-Native Solution for HPC-Based Services (2024). <https://arxiv.org/abs/2407.00110>
- [251] Wang, Y., Chen, K., Tan, H., Guo, K.: Tabi: An efficient multi-level inference system for large language models. In: Proceedings of the Eighteenth European Conference on Computer Systems, pp. 233–248 (2023)
- [252] Peng, Z., Wang, Z., Deng, D.: Near-duplicate sequence search at scale for large language model memorization evaluation. *Proceedings of the ACM on Management of Data* **1**(2), 1–18 (2023)

- [253] Miao, X., Shi, C., Duan, J., Xi, X., Lin, D., Cui, B., Jia, Z.: Spotserve: Serving generative large language models on preemptible instances. arXiv preprint arXiv:2311.15566 (2023)
- [254] Huang, K., Yin, H., Huang, H., Gao, W.: Towards green ai in fine-tuning large language models via adaptive backpropagation. arXiv preprint arXiv:2309.13192 (2023)
- [255] Jiao, X., Yin, Y., Shang, L., Jiang, X., Chen, X., Li, L., Wang, F., Liu, Q.: Tinybert: Distilling bert for natural language understanding. arXiv preprint arXiv:1909.10351 (2019)
- [256] Shen, H., Hanwen, C., Bo, D., Yu, L., Hengyu, M.: Efficient llm inference on cpus. arXiv preprint arXiv:2311.00502 (2023)
- [257] Wu, Y., Zhao, Y., Hu, B., Minervini, P., Stenetorp, P., Riedel, S.: An efficient memory-augmented transformer for knowledge-intensive nlp tasks. arXiv preprint arXiv:2210.16773 (2022)
- [258] Ma, Z., Ethayarajh, K., Thrush, T., Jain, S., Wu, L., Jia, R., Potts, C., Williams, A., Kiela, D.: Dynaboard: An evaluation-as-a-service platform for holistic next-generation benchmarking. Advances in Neural Information Processing Systems **34**, 10351–10367 (2021)
- [259] Strubell, E., Ganesh, A., McCallum, A.: Energy and policy considerations for deep learning in nlp. arXiv preprint arXiv:1906.02243 (2019)
- [260] Schmidt, V., Goyal, K., Joshi, A., Feld, B., Conell, L., Laskaris, N., Blank, D., Wilson, J., Friedler, S., Luccioni, S.: Codecarbon: estimate and track carbon emissions from machine learning computing. Cited on, 20 (2021)
- [261] Anthony, L.F.W., Kanding, B., Selvan, R.: Carbontracker: Tracking and predicting the carbon footprint of training deep learning models. arXiv preprint arXiv:2007.03051 (2020)
- [262] Henderson, P., Hu, J., Romoff, J., Brunskill, E., Jurafsky, D., Pineau, J.: Towards the systematic reporting of the energy and carbon footprints of machine learning. The Journal of Machine Learning Research **21**(1), 10039–10081 (2020)
- [263] Lacoste, A., Luccioni, A., Schmidt, V., Dandres, T.: Quantifying the carbon emissions of machine learning. arXiv preprint arXiv:1910.09700 (2019)
- [264] Faiz, A., Kaneda, S., Wang, R., Osi, R., Sharma, P., Chen, F., Jiang, L.: Llmcarbon: Modeling the end-to-end carbon footprint of large language models. arXiv preprint arXiv:2309.14393 (2023)
- [265] Wu, C., Zhang, H., Ju, L., Huang, J., Xiao, Y., Huan, Z., Li, S., Meng, F., Liang,

- L., Zhang, X., et al.: Rethinking memory and communication cost for efficient large language model training. arXiv preprint arXiv:2310.06003 (2023)
- [266] Xu, C., Zhou, W., Ge, T., Xu, K., McAuley, J., Wei, F.: Beyond preserved accuracy: Evaluating loyalty and robustness of bert compression. arXiv preprint arXiv:2109.03228 (2021)
- [267] Stanton, S., Izmailov, P., Kirichenko, P., Alemi, A.A., Wilson, A.G.: Does knowledge distillation really work? Advances in Neural Information Processing Systems **34**, 6906–6919 (2021)
- [268] Hooker, S., Moorosi, N., Clark, G., Bengio, S., Denton, E.: Characterising bias in compressed models. arXiv preprint arXiv:2010.03058 (2020)
- [269] Su, D., Zhang, H., Chen, H., Yi, J., Chen, P.-Y., Gao, Y.: Is robustness the cost of accuracy?—a comprehensive study on the robustness of 18 deep image classification models. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 631–648 (2018)
- [270] Treviso, M., Lee, J.-U., Ji, T., Aken, B.v., Cao, Q., Ciosici, M.R., Hassid, M., Heafield, K., Hooker, S., Raffel, C., et al.: Efficient methods for natural language processing: A survey. Transactions of the Association for Computational Linguistics **11**, 826–860 (2023)
- [271] Pareto, V.: Cours D'économie Politique vol. 1. Librairie Droz, ??? (1964)
- [272] Santhanam, K., Saad-Falcon, J., Franz, M., Khattab, O., Sil, A., Florian, R., Sultan, M.A., Roukos, S., Zaharia, M., Potts, C.: Moving beyond downstream task accuracy for information retrieval benchmarking. arXiv preprint arXiv:2212.01340 (2022)
- [273] Liu, X., Sun, T., He, J., Wu, J., Wu, L., Zhang, X., Jiang, H., Cao, Z., Huang, X., Qiu, X.: Towards efficient nlp: A standard evaluation and a strong baseline. arXiv preprint arXiv:2110.07038 (2021)
- [274] Naveed, H., Khan, A.U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Barnes, N., Mian, A.: A comprehensive overview of large language models. arXiv preprint arXiv:2307.06435 (2023)
- [275] Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., Bowman, S.R.: Glue: A multi-task benchmark and analysis platform for natural language understanding. arXiv preprint arXiv:1804.07461 (2018)
- [276] Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., Bowman, S.: Superglue: A stickier benchmark for general-purpose language understanding systems. Advances in neural information processing systems **32** (2019)

- [277] Bojar, O., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huck, M., Yepes, A.J., Koehn, P., Logacheva, V., Monz, C., *et al.*: Findings of the 2016 conference on machine translation (wmt16). In: First Conference on Machine Translation, pp. 131–198 (2016). Association for Computational Linguistics
- [278] Barrault, L., Biesialska, M., Bojar, O., Costa-jussà, M.R., Federmann, C., Graham, Y., Grundkiewicz, R., Haddow, B., Huck, M., Joanis, E., *et al.*: Findings of the 2020 conference on machine translation (wmt20). In: Proceedings of the Fifth Conference on Machine Translation, pp. 1–55 (2020)
- [279] Rajpurkar, P., Zhang, J., Lopyrev, K., Liang, P.: Squad: 100,000+ questions for machine comprehension of text. arXiv preprint arXiv:1606.05250 (2016)
- [280] Rajpurkar, P., Jia, R., Liang, P.: Know what you don’t know: Unanswerable questions for squad. arXiv preprint arXiv:1806.03822 (2018)
- [281] Min, S., Boyd-Graber, J., Alberti, C., Chen, D., Choi, E., Collins, M., Guu, K., Hajishirzi, H., Lee, K., Palomaki, J., *et al.*: Neurips 2020 efficientqa competition: Systems, analyses and lessons learned. In: NeurIPS 2020 Competition and Demonstration Track, pp. 86–111 (2021). PMLR
- [282] Wang, A., Wolf, T.: Overview of the sustainlp 2020 shared task. In: Proceedings of SustaiNLP: Workshop on Simple and Efficient Natural Language Processing, pp. 174–178 (2020)
- [283] Zhou, W., Zeng, Y., Diao, S., Zhang, X.: Vlue: A multi-task benchmark for evaluating vision-language models. arXiv preprint arXiv:2205.15237 (2022)
- [284] Tay, Y., Dehghani, M., Abnar, S., Shen, Y., Bahri, D., Pham, P., Rao, J., Yang, L., Ruder, S., Metzler, D.: Long range arena: A benchmark for efficient transformers. arXiv preprint arXiv:2011.04006 (2020)
- [285] Bajaj, P., Campos, D., Craswell, N., Deng, L., Gao, J., Liu, X., Majumder, R., McNamara, A., Mitra, B., Nguyen, T., *et al.*: Ms marco: A human generated machine reading comprehension dataset. arXiv preprint arXiv:1611.09268 (2016)
- [286] Ford, B.W., Zong, Z.: Portauthority: Integrating energy efficiency analysis into cross-platform development cycles via dynamic program analysis. Sustainable Computing: Informatics and Systems **30**, 100530 (2021)
- [287] Xie, Y., Chen, S., Ni, Q., Wu, H.: Integration of resource allocation and task assignment for optimizing the cost and maximum throughput of business processes. Journal of Intelligent Manufacturing **30**(3), 1351–1369 (2019)
- [288] Shamshirband, S., Joloudari, J.H., Shirkharkolaie, S.K., Mojrian, S., Rahmani, F., Mostafavi, S., Mansor, Z.: Game theory and evolutionary optimization

approaches applied to resource allocation problems in computing environments: A survey. Mathematical Biosciences and Engineering (2021)

- [289] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Chi, E., Le, Q., Zhou, D.: Chain of thought prompting elicits reasoning in large language models. arXiv preprint arXiv:2201.11903 (2022)
- [290] Diao, S., Wang, P., Lin, Y., Zhang, T.: Active prompting with chain-of-thought for large language models. arXiv preprint arXiv:2302.12246 (2023)
- [291] Vilalta, R., Drissi, Y.: A perspective view and survey of meta-learning. Artificial intelligence review **18**, 77–95 (2002)
- [292] Elsken, T., Metzen, J.H., Hutter, F.: Neural architecture search: A survey. The Journal of Machine Learning Research **20**(1), 1997–2017 (2019)
- [293] Murshed, M.S., Murphy, C., Hou, D., Khan, N., Ananthanarayanan, G., Hussain, F.: Machine learning at the network edge: A survey. ACM Computing Surveys (CSUR) **54**(8), 1–37 (2021)
- [294] Shi, W., Cao, J., Zhang, Q., Li, Y., Xu, L.: Edge computing: Vision and challenges. IEEE internet of things journal **3**(5), 637–646 (2016)
- [295] Zhang, C., Xie, Y., Bai, H., Yu, B., Li, W., Gao, Y.: A survey on federated learning. Knowledge-Based Systems **216**, 106775 (2021)