Preprints.org

Review

Not peer-reviewed version

Exploring the Landscape of Large and Small Language Models: Advancements, Trade-offs, and Future Directions

Duha Shams , Ikraam Salama , Idowu Callixtus *

Posted Date: 7 January 2025

doi: 10.20944/preprints202501.0502.v1

Keywords: Large Language Models; Small Language Models; Natural Language Processing; Model Compression; Knowledge Distillation; Model Pruning; Computational Efficiency; Task Specialization; Resource Constraints; Real-time Applications; Energy Efficiency; Hybrid Approaches; Deep Learning; NLP Models; Sustainability in AI

Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 January 2025 doi:10.20944/preprints202501.0502.v1

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Review

Exploring the Landscape of Large and Small Language Models: Advancements, Trade-offs, and Future Directions

Duha Shams †**, Ikraam Salama** † **and Callixtus Idowu** *

KAUST, King Abdullah University of Science and Technology, Arabie, Saoudite, 23955, Thuwal, Kingdom of Saudi Arabia

* Correspondence: callixtus.idowu@kaust.edu.sa

† These authors contributed equally to this work.

Abstract: Recent advances in natural language processing (NLP) have led to the development of large language models (LLMs) and small language models (small LMs), which have revolutionized the field. LLMs, such as GPT-3 and PaLM, are capable of performing a wide range of tasks with state-of-the-art accuracy, thanks to their vast number of parameters and extensive training data. However, these models are resource-intensive, requiring significant computational power for both training and deployment. In contrast, small LMs offer a more efficient alternative, with reduced computational requirements and faster inference times, making them well-suited for resource-constrained environments such as mobile devices and real-time applications. This survey explores the key differences between LLMs and small LMs, focusing on aspects such as model size, computational efficiency, performance, and deployment scenarios. We also discuss the trade-offs associated with selecting between the two, and highlight techniques such as knowledge distillation and model pruning that are used to optimize small LMs. Finally, we examine the future directions of language model research, including hybrid approaches that combine the strengths of both LLMs and small LMs, and advancements aimed at improving energy efficiency and sustainability. Our goal is to provide a comprehensive overview of the current landscape of LLMs and small LMs, and to offer insights into the ongoing challenges and opportunities in the field of NLP.

Keywords: large language models; small language models; natural language processing; model compression; knowledge distillation; model pruning; computational efficiency; task specialization; resource constraints; real-time applications; energy efficiency; hybrid approaches; deep learning; NLP models; sustainability in AI

1. Introduction

Large Language Models (LLMs) have emerged as one of the most transformative technologies in the field of artificial intelligence (AI) and natural language processing (NLP). These models, exemplified by architectures such as OpenAI's GPT series, BERT, and T5, are capable of processing and generating human-like text with remarkable accuracy and fluency. They have found applications in a wide range of domains, from customer service chatbots and automated content generation to complex scientific research and creative writing. The capabilities of LLMs are underpinned by their size, which often involves billions, or even trillions, of parameters trained on extensive and diverse datasets. This scale enables LLMs to capture intricate patterns in language, making them versatile tools for a variety of tasks. However, the impressive performance of LLMs comes at a cost [1]. The computational resources required to train, fine-tune, and deploy these models are immense, often necessitating specialized hardware such as GPUs or TPUs, and significant energy consumption. For example, the training process of models like GPT-3 involves several petaflop-days of compute, translating into substantial financial and environmental costs. Moreover, the deployment of these models requires high memory bandwidth and storage capacity, making them inaccessible for many researchers, developers, and

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

organizations, particularly those operating in resource-constrained environments [2]. The reliance on extensive computational resources has also raised ethical concerns, including the carbon footprint of training large models and the equitable distribution of AI capabilities [3]. In addition to resource concerns, LLMs face challenges related to latency and real-time performance [4]. Their size and complexity often result in slower inference times, which can be a critical limitation in applications requiring instant responses, such as voice assistants or real-time translation systems. Furthermore, the dependency on high-end infrastructure restricts the use of LLMs in settings such as mobile devices, edge computing environments, and low-power IoT devices, where computational and energy efficiency are paramount [5]. The transition from Large Language Models (LLMs) to Small Language Models (SLMs) represents a strategic response to these challenges. SLMs are designed to offer similar capabilities to LLMs while being significantly more compact and efficient [6]. This is achieved through a variety of techniques, including model compression, knowledge distillation, quantization, and pruning. By leveraging these approaches, researchers aim to create models that retain the essential functionalities of LLMs while being optimized for deployment in resource-constrained settings. The shift towards SLMs is not merely a technical adjustment but also an enabler of broader access and inclusivity in AI technologies [7]. Smaller models lower the barrier to entry for organizations and individuals who may not have access to state-of-the-art infrastructure, democratizing the benefits of NLP advancements. Moreover, SLMs are better suited for deployment in applications tailored to specific domains or languages, particularly those underrepresented in mainstream AI research. For example, SLMs can be fine-tuned and deployed for regional languages, enhancing their utility in diverse cultural and linguistic contexts. This survey aims to provide a comprehensive overview of the journey from LLMs to SLMs, exploring the motivations, techniques, and applications that define this transition [8]. We begin by examining the foundational aspects of LLMs, including their architecture, training paradigms, and performance characteristics. Next, we delve into the limitations of LLMs, focusing on the computational, environmental, and accessibility challenges they present. Following this, we explore the methodologies employed to create SLMs, highlighting key advances in model compression, knowledge distillation, and related areas [9]. Finally, we discuss the practical implications and future directions of SLM research, emphasizing their potential to redefine the landscape of NLP and AI more broadly [10]. By addressing the transition from LLMs to SLMs, this survey seeks to contribute to the ongoing dialogue on sustainable and inclusive AI development [11]. It is our hope that this work will inspire further innovation and collaboration in the quest to make NLP technologies more efficient, accessible, and impactful.

1.1. Scope and Objectives

The scope of this survey encompasses the transition from Large Language Models (LLMs) to Small Language Models (SLMs), a field that addresses critical challenges in computational efficiency, accessibility, and environmental sustainability within AI and NLP [12]. This work is intended to serve as a foundational reference for researchers, practitioners, and decision-makers interested in understanding and contributing to this rapidly evolving domain. The primary objectives of this survey are:

- To review methods for compressing LLMs into SLMs, including state-of-the-art techniques such as quantization, pruning, knowledge distillation, and low-rank factorization. By examining these methods, we aim to provide a comprehensive understanding of how model compression is achieved without compromising functionality.
- To evaluate trade-offs between model size and performance, focusing on critical metrics such as accuracy, latency, memory usage, and energy efficiency. This evaluation includes an analysis of the contexts in which these trade-offs are most significant, such as edge computing, mobile devices, and domain-specific applications.
- To discuss the practical deployment of SLMs in real-world scenarios, including their application in under-resourced languages, low-power IoT environments, and industries with stringent

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 January 2025 doi:10.20944/preprints202501.0502.v1

computational constraints. This objective also encompasses an exploration of case studies where SLMs have successfully addressed specific challenges.

- To identify gaps and opportunities in current research, highlighting areas where further innovation is needed. This includes the development of automated tools for transitioning LLMs to SLMs and novel approaches to enhance the interpretability and fairness of smaller models.
- To foster a dialogue on ethical considerations, such as the environmental impact of model training and deployment, and the equitable distribution of AI technologies facilitated by SLMs [13]. By addressing these issues, we aim to align the development of SLMs with broader societal goals [14].

In achieving these objectives, this survey seeks to bridge the gap between theoretical advancements and practical implementations, promoting a holistic approach to the design and adoption of Small Language Models.

2. Background

The evolution of natural language processing (NLP) has been a journey marked by continuous improvements in machine learning techniques. From early symbolic models to the latest transformer-based deep learning architectures, the development of language models has been driven by advances in computational power, data availability, and algorithmic innovations [15]. This section provides an overview of the key milestones in the development of language models, with a focus on the transition from early statistical models to modern large and small language models (LLMs and small LMs).

2.1. Early Language Models and Statistical Approaches

The field of NLP began with rule-based systems and symbolic models, which relied on linguistic rules and dictionaries to process text. These early systems, while effective in specific domains, struggled with ambiguity and the complexity of natural language. In the 1980s and 1990s, researchers began to explore statistical models that could learn patterns from data, paving the way for probabilistic approaches to language processing. The most prominent of these early statistical models were n-gram models, which estimated the probability of a word given the previous $n-1$ words in a sequence. While simple and effective for certain tasks such as speech recognition and machine translation, n-gram models had limitations in capturing long-range dependencies and semantic relationships. They also required large amounts of training data to achieve reliable performance. The introduction of machine learning algorithms such as decision trees, support vector machines, and hidden Markov models (HMMs) further advanced the field, allowing for more robust models capable of handling a variety of NLP tasks, such as part-of-speech tagging and named entity recognition.

2.2. Neural Networks and Word Embeddings

In the 2000s, neural networks began to gain prominence in NLP. Early work focused on applying deep learning techniques to tasks such as document classification and sentiment analysis. One of the key breakthroughs during this period was the development of word embeddings, particularly models like Word2Vec and GloVe, which represented words as dense vectors in a continuous vector space. These embeddings allowed words with similar meanings to be mapped to nearby points in the vector space, capturing semantic relationships like synonyms and analogies [16]. Word embeddings, however, were limited in their ability to represent word meanings in context. They represented words as static vectors, independent of the surrounding words or sentence structure. This limitation prompted the development of more advanced models, such as recurrent neural networks (RNNs) and Long Short-Term Memory (LSTM) networks, which were capable of processing sequences of words and capturing context-sensitive word meanings.

2.3. The Rise of Transformer Models

The introduction of the Transformer architecture by Vaswani et al [17]. in 2017 marked a pivotal moment in the field of NLP. The Transformer model eliminated the need for recurrence in sequence

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 January 2025 doi:10.20944/preprints202501.0502.v1

processing, instead using a self-attention mechanism to allow the model to attend to all words in a sequence simultaneously. This innovation drastically improved the efficiency and scalability of NLP models, as it allowed for parallelization during training and better handling of long-range dependencies. The Transformer model was the foundation for many subsequent advancements in NLP. It enabled the development of large-scale pre-trained language models, which could be fine-tuned for a variety of downstream tasks. These models included OpenAI's GPT series, Google's BERT, and T5, among others. By pre-training on massive amounts of text data and fine-tuning on specific tasks, these models achieved state-of-the-art performance across a wide range of applications, including machine translation, text generation, and question answering. The success of Transformer-based models also sparked the development of even larger models, such as GPT-3, which contains 175 billion parameters [18]. These models demonstrated impressive performance on tasks requiring deep reasoning, creativity, and even understanding of context beyond the scope of traditional NLP systems.

2.4. The Emergence of Large Language Models (LLMs)

Large language models (LLMs) represent the next step in the evolution of NLP. These models, with their vast number of parameters, are able to generalize across a wide range of NLP tasks without task-specific tuning. LLMs like GPT-3 and Google's PaLM leverage their scale to capture nuanced patterns in language and exhibit remarkable capabilities in areas such as text generation, summarization, question answering, and even creative writing [19]. The primary strength of LLMs lies in their ability to learn from vast corpora of text data and generalize to a variety of tasks without requiring explicit task-specific training data [20]. This generalization is achieved through pre-training on massive datasets followed by fine-tuning for specific tasks [21]. The advent of LLMs has led to significant advancements in fields such as conversational AI, document generation, and automated content creation. However, despite their impressive performance, LLMs are not without challenges [22]. They require substantial computational resources to train, making them expensive to develop and deploy. The environmental impact of training these models has also raised concerns, as the energy consumption associated with training large models is substantial. Furthermore, LLMs tend to exhibit biases learned from the data they were trained on, which can lead to undesirable outcomes when applied to real-world scenarios.

2.5. The Emergence of Small Language Models (Small LMs)

As the demand for more efficient models grew, researchers began to focus on creating smaller, more computationally efficient language models that could deliver high performance without the heavy resource requirements of LLMs [23]. Small language models (small LMs) are designed to retain much of the power of large models while reducing their size and computational cost. The development of small LMs has been facilitated by several key techniques, including knowledge distillation, model pruning, and quantization [24]. Knowledge distillation, for example, involves training a smaller model (the student) to mimic the behavior of a larger, pre-trained model (the teacher) [25]. This allows the smaller model to approximate the performance of the larger model while requiring fewer parameters and less computation. DistilBERT, a smaller version of BERT, is one such example where knowledge distillation has been successfully applied. Other methods, such as pruning and quantization, further reduce the size of language models by eliminating redundant weights or representing model parameters with lower precision. These techniques have made it possible to deploy powerful language models in environments with limited resources, such as mobile devices, embedded systems, and real-time applications. Small LMs are particularly valuable in scenarios where real-time inference is required, such as virtual assistants, chatbots, and speech recognition systems. They are also more energy-efficient, which is critical for sustainable AI applications.

2.6. Challenges and Trade-Offs in Model Selection

While large language models offer state-of-the-art performance across a wide range of tasks, they come with significant trade-offs in terms of computational cost, memory requirements, and environmental impact [26]. In contrast, small language models offer efficiency and faster inference

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 January 2025 doi:10.20944/preprints202501.0502.v1

times but may not achieve the same level of performance on certain complex tasks [27]. The choice between large and small language models depends on the specific requirements of the application, such as the trade-off between accuracy and resource usage, as well as the deployment environment. In some cases, hybrid approaches may be used, where a large model is used for initial processing or generation, and a smaller model is deployed for specific tasks or real-time inference [28]. This approach aims to combine the strengths of both large and small models, optimizing both performance and efficiency [29].

2.7. Current Landscape and Future Directions

The landscape of language models is rapidly evolving, with ongoing research focused on improving both large and small models [30]. For large models, efforts are being made to improve training efficiency, reduce energy consumption, and mitigate biases [31]. For small models, researchers are exploring new methods for maintaining high performance while further reducing model size and computational requirements. Recent trends include the development of domain-specific models that are smaller but highly optimized for particular tasks or industries. Furthermore, there is growing interest in multilingual models that can handle multiple languages without requiring separate models for each language [32]. The future of language models is likely to see a balance between the development of large-scale, general-purpose models and smaller, task-specific models [33]. Researchers are also exploring the potential of "edge AI," where small models are deployed on devices such as smartphones, wearables, and IoT devices to process language data locally, reducing reliance on cloud-based computation and improving privacy. In the next section, we will delve into the differences between large language models and small language models, exploring their respective advantages, challenges, and use cases in modern NLP applications [34].

3. Key Differences Between Large Language Models (LLMs) and Small Language Models (Small LMs)

In this section, we explore the key differences between large language models (LLMs) and small language models (small LMs) [35]. These differences stem from several factors, including model size, computational requirements, training data, performance characteristics, and use cases [36]. Understanding these differences is essential for selecting the appropriate model for a given NLP task or deployment scenario [37].

3.1. Model Size and Parameters

The most obvious difference between LLMs and small LMs is their size. LLMs, such as GPT-3 and PaLM, consist of billions or even trillions of parameters [38]. These models are trained on massive datasets that span a wide variety of domains, allowing them to learn a diverse set of patterns in language. The large number of parameters enables LLMs to capture complex relationships and generate highly coherent, contextually relevant outputs across multiple tasks [39]. In contrast, small LMs are designed with fewer parameters, typically ranging from a few million to a few hundred million. By reducing the number of parameters, these models are less resource-intensive and more suitable for environments with limited computational power, such as mobile devices or edge computing platforms [40]. Despite their smaller size, small LMs often perform surprisingly well on specific tasks, especially when fine-tuned on domain-specific datasets [41].

3.2. Computational Requirements

The computational requirements for training and deploying LLMs are significantly higher than for small LMs. LLMs require specialized hardware, such as high-performance GPUs or TPUs, to handle the vast amount of data and parameters during both training and inference [42]. Training an LLM can take weeks or even months, depending on the scale of the model and the hardware resources available. Additionally, inference with LLMs can be slow and resource-intensive, making them less suitable for real-time applications where low latency is critical [43]. Small LMs, on the other hand, are optimized for efficiency [44]. They require less memory, computational power, and storage, which makes them more feasible for deployment in resource-constrained environments. Training small LMs is faster and less expensive, and their inference times are much quicker, making them ideal for real-time applications such as chatbots, speech recognition, or text classification.

3.3. Training Data and Generalization

One of the key advantages of LLMs is their ability to generalize across a wide range of tasks [45]. This generalization is made possible by the large amounts of diverse training data used to pre-train LLMs. These models are typically trained on vast corpora of text data, including books, articles, websites, and other publicly available documents. The broad coverage of topics and domains enables LLMs to perform well on a variety of NLP tasks, including text generation, question answering, summarization, and more, without the need for extensive task-specific training data. Small LMs, by contrast, are often trained on smaller, more specialized datasets. This limitation can restrict their ability to generalize across a wide range of tasks, but small LMs can still perform exceptionally well in narrow, domain-specific areas. For example, small LMs may be fine-tuned for particular industries or applications, such as legal document analysis or medical text processing, where they can outperform larger models on specialized tasks despite their smaller size.

3.4. Performance and Accuracy

LLMs generally outperform small LMs in terms of raw performance across a variety of NLP benchmarks [46]. The large number of parameters and the extensive training data enable LLMs to capture nuanced patterns and relationships in language, which translates into higher accuracy on complex tasks. LLMs also exhibit better generalization when faced with novel or unseen tasks, as they can leverage their broad understanding of language. However, small LMs can still achieve competitive performance on specific tasks, especially when they are fine-tuned for a particular application or domain. In some cases, small LMs may even outperform larger models on certain tasks if they are better optimized for those tasks or if they are trained on high-quality, domain-specific data [47]. Small LMs tend to be more lightweight and faster at inference, which can be a crucial advantage in real-time applications [48].

3.5. Energy Efficiency and Sustainability

The energy consumption of training and deploying large language models is a growing concern, as the environmental impact of these models is significant [49]. LLMs require enormous amounts of energy to train, and even after training, running them in production can consume substantial computational resources. As a result, there is increasing pressure to make AI models more sustainable and energy-efficient, which has led to the exploration of techniques like model pruning, distillation, and quantization. Small LMs are far more energy-efficient than their larger counterparts. Because they are smaller in size and require fewer parameters, small LMs consume less power both during training and inference [50]. This makes them more sustainable and suitable for deployment in scenarios where energy consumption is a key concern, such as in mobile devices, IoT applications, or low-latency environments. By using less power, small LMs can help reduce the carbon footprint of AI deployments [51].

3.6. Deployment Scenarios and Use Cases

The choice between LLMs and small LMs largely depends on the deployment scenario and the specific requirements of the application. LLMs are best suited for large-scale applications where high performance and versatility are required [52]. They are used in tasks that involve complex language generation, such as content creation, chatbots, virtual assistants, and large-scale document processing. Due to their ability to generalize across tasks, LLMs are also well-suited for tasks that require reasoning, such as answering complex questions or engaging in long-form conversations [53]. On the other hand, small LMs are ideal for resource-constrained environments where real-time processing is necessary [54]. These models are commonly deployed in mobile applications, such as

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 January 2025 doi:10.20944/preprints202501.0502.v1

speech recognition on smartphones, real-time chatbots, or predictive text input [55]. Small LMs are also deployed in embedded systems, where low latency and high efficiency are essential [56]. In industries like healthcare, finance, and legal services, small LMs are fine-tuned for specific tasks, such as medical diagnosis assistance, financial forecasting, or legal document analysis. Small LMs are also becoming increasingly important in edge AI applications, where data is processed locally on devices such as smart speakers, wearables, or autonomous vehicles. In these cases, small LMs can operate in environments with limited bandwidth and computational resources, providing fast, on-device responses while maintaining privacy and security by keeping data processing local.

3.7. Trade-Offs and Hybrid Approaches

When selecting between LLMs and small LMs, organizations must carefully consider the trade-offs between performance, cost, and resource requirements. LLMs offer state-of-the-art performance but come with high computational costs, long training times, and challenges related to deployment and sustainability [57]. Small LMs, while more efficient, may not perform as well on complex or general tasks but are highly suitable for domain-specific applications that require fast, efficient inference. In many cases, hybrid approaches can be used to combine the strengths of both LLMs and small LMs. For example, a large model might be used for pre-processing or initial generation, and then a smaller model could be deployed for specific tasks or real-time responses [58]. This approach can help balance the trade-offs between accuracy and efficiency, allowing for high-performance applications that are also cost-effective and sustainable.

3.8. Future Directions and Ongoing Research

The future of language models lies in improving both large and small models. For LLMs, research is focused on making these models more efficient by developing techniques to reduce their computational cost, such as efficient transformers, multi-modal learning, and fine-tuning strategies. For small LMs, ongoing work is focused on developing methods to improve their performance without increasing their size, such as through task-specific fine-tuning, knowledge transfer, and domain adaptation. There is also growing interest in multi-modal models that can handle both text and other forms of data, such as images, audio, and video [59]. These models will require both large and small LMs to work in tandem to process diverse types of information and perform more complex tasks [60]. As the field of NLP continues to advance, the development of more efficient and specialized models will continue to push the boundaries of what is possible with language technology [61]. In the next section, we will explore the advantages and challenges associated with both LLMs and small LMs in more detail, examining how each model type is applied across various domains and use cases.

4. Conclusion

In this survey, we have explored the evolution and key characteristics of large language models (LLMs) and small language models (small LMs), emphasizing the differences between them in terms of size, computational requirements, training data, performance, and deployment scenarios. Both LLMs and small LMs have their unique strengths and challenges, and the choice between them depends largely on the specific needs of the task at hand. LLMs, with their vast number of parameters and extensive training data, are capable of delivering state-of-the-art performance across a wide range of natural language processing (NLP) tasks [62]. Their ability to generalize across multiple domains makes them highly versatile, excelling in applications that require complex language understanding, reasoning, and generation. However, their computational and memory requirements present significant challenges in terms of both training and deployment, leading to concerns about resource consumption, energy efficiency, and environmental sustainability [63]. On the other hand, small LMs offer a more efficient alternative, with reduced computational overhead and faster inference times. These models are highly suitable for resource-constrained environments such as mobile devices, edge computing, and real-time applications where low latency is crucial. While they may not achieve the same level of performance as LLMs in more complex tasks, small LMs can still perform exceptionally well when fine-tuned on domain-specific data. They are particularly valuable in scenarios where task specialization, energy efficiency, and rapid response times are prioritized. The rapid advancement in techniques for model compression, such as knowledge distillation, quantization, and pruning, has enabled the development of small LMs that can approximate the performance of larger models while significantly reducing their resource requirements. This has opened up new possibilities for deploying high-performance language models in a variety of applications, from mobile applications to IoT devices, without compromising on efficiency or speed. Looking forward, the future of language models is likely to involve a hybrid approach, where large models and small models are used in tandem to leverage the strengths of both. As research continues to advance, we expect to see further innovations that optimize the efficiency of large models, improve the accuracy of small models, and explore new multi-modal applications that combine language processing with other forms of data, such as images and audio. In conclusion, both LLMs and small LMs play critical roles in the landscape of modern NLP, each serving distinct use cases and deployment environments. As the field continues to evolve, researchers will need to carefully balance the trade-offs between accuracy, efficiency, and sustainability, tailoring model selection to the specific needs of each application [64]. The ongoing development of more efficient algorithms, model architectures, and training techniques will be crucial in making these models more accessible, affordable, and environmentally sustainable for a broader range of real-world applications [65].

References

1. Dettmers, T., Lewis, M., Belkada, Y., Zettlemoyer, L.: Llm. int8 (): 8-bit matrix multiplication for transformers at scale. arXiv preprint arXiv:2208.07339 (2022)
2. Wen, Y., Li, Z., Du, W., Mou, L.: f-Divergence Minimization for Sequence-Level Knowledge Distillation (2023). https://arxiv.org/abs/2307.15190
3. Zniyed, Y., Nguyen, T.P., *et al.*: Efficient tensor decomposition-based filter pruning. Neural Networks **178**, 106393 (2024)
4. Ainslie, J., Lee-Thorp, J., Jong, M., Zemlyanskiy, Y., Lebron, F., Sanghai, S.: GQA: Training generalized multi-query transformer models from multi-head checkpoints. In: EMNLP (2023)
5. Computer, T.: RedPajama: An Open Source Recipe to Reproduce LLaMA Training Dataset. https://github.com/togethercomputer/RedPajama-Data
6. Stojkovic, J., Choukse, E., Zhang, C., Goiri, I., Torrellas, J.: Towards Greener LLMs: Bringing Energy-Efficiency to the Forefront of LLM Inference (2024). https://arxiv.org/abs/2403.20306
7. Ge, S., Zhang, Y., Liu, L., Zhang, M., Han, J., Gao, J.: Model tells you what to discard: Adaptive KV cache compression for LLMs. In: The Twelfth International Conference on Learning Representations (2024). https://openreview.net/forum?id=uNrFpDPMyo
8. Hu, J., Ruder, S., Siddhant, A., Neubig, G., Firat, O., Johnson, M.: Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In: International Conference on Machine Learning, pp. 4411–4421 (2020). PMLR
9. Li, Q., Hong, J., Xie, C., Tan, J., Xin, R., Hou, J., Yin, X., Wang, Z., Hendrycks, D., Wang, Z., Li, B., He, B., Song, D.: LLM-PBE: Assessing Data Privacy in Large Language Models (2024). https://arxiv.org/abs/2408.12787
10. Kitaev, N., Kaiser, Ł., Levskaya, A.: Reformer: The efficient transformer. arXiv preprint arXiv:2001.04451 (2020)
11. Mirzadeh, S.I., Alizadeh-Vahid, K., Mehta, S., Mundo, C.C., Tuzel, O., Samei, G., Rastegari, M., Farajtabar, M.: ReLU strikes back: Exploiting activation sparsity in large language models. In: The Twelfth International Conference on Learning Representations (2024). https://openreview.net/forum?id=osoWxY8q2E
12. 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
13. Papineni, K., Roukos, S., Ward, T., Zhu, W.-J.: BLEU: a Method for Automatic Evaluation of Machine Translation (2002). https://www.aclweb.org/anthology/W02-2019.pdf
14. 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)

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 January 2025 doi:10.20944/preprints202501.0502.v1

15. Liu, Y.: Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
16. Liu, Y., Deng, G., Li, Y., Wang, K., Wang, Z., Wang, X., Zhang, T., Liu, Y., Wang, H., Zheng, Y., Liu, Y.: Prompt Injection attack against LLM-integrated Applications (2024). https://arxiv.org/abs/2306.05499
17. Boizard, N., Haddad, K.E., Hudelot, C., Colombo, P.: Towards Cross-Tokenizer Distillation: the Universal Logit Distillation Loss for LLMs (2024). https://arxiv.org/abs/2402.12030
18. Liu, J., Gong, R., Wei, X., Dong, Z., Cai, J., Zhuang, B.: Qllm: Accurate and efficient low-bitwidth quantization for large language models. arXiv preprint arXiv:2310.08041 (2023)
19. Abdin, M., Aneja, J., Awadalla, H., et al.: Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone (2024). https://arxiv.org/abs/2404.14219
20. Utama, P.A., Moosavi, N.S., Gurevych, I.: Towards debiasing NLU models from unknown biases. In: Webber, B., Cohn, T., He, Y., Liu, Y. (eds.) Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 7597–7610. Association for Computational Linguistics, Online (2020). https://doi.org/10.18653/v1/2020.emnlp-main.613 . https://aclanthology.org/2020.emnlp-main.613
21. Lu, Z., Li, X., Cai, D., Yi, R., Liu, F., Zhang, X., Lane, N.D., Xu, M.: Small language models: Survey, measurements, and insights. arXiv preprint arXiv:2409.15790 (2024)
22. Santacroce, M., Wen, Z., Shen, Y., Li, Y.: What matters in the structured pruning of generative language models? arXiv preprint arXiv:2302.03773 (2023)
23. Xu, J., Li, Z., Chen, W., Wang, Q., Gao, X., Cai, Q., Ling, Z.: On-Device Language Models: A Comprehensive Review (2024). https://arxiv.org/abs/2409.00088
24. Zniyed, Y., Nguyen, T.P., et al.: Enhanced network compression through tensor decompositions and pruning. IEEE Transactions on Neural Networks and Learning Systems (2024)
25. Sarkar, R., Liang, H., Fan, Z., Wang, Z., Hao, C.: Edge-MoE: Memory-Efficient Multi-Task Vision Transformer Architecture with Task-level Sparsity via Mixture-of-Experts (2023). https://arxiv.org/abs/2305.18691
26. Yang, S., Ali, M.A., Wang, C.-L., Hu, L., Wang, D.: Moral: Moe augmented lora for llms' lifelong learning. arXiv preprint arXiv:2402.11260 (2024)
27. Jawahar, G., Yang, H., Xiong, Y., Liu, Z., Wang, D., Sun, F., Li, M., Pappu, A., Oguz, B., Abdul-Mageed, M., et al.: Mixture-of-supernets: Improving weight-sharing supernet training with architecture-routed mixture-of-experts. arXiv preprint arXiv:2306.04845 (2023)
28. Ma, S., Wang, H., Ma, L., Wang, L., Wang, W., Huang, S., Dong, L., Wang, R., Xue, J., Wei, F.: The era of 1-bit llms: All large language models are in 1.58 bits. arXiv preprint arXiv:2402.17764 (2024)
29. Dettmers, T., Lewis, M., Belkada, Y., Zettlemoyer, L.: Gpt3. int8 (): 8-bit matrix multiplication for transformers at scale. Advances in Neural Information Processing Systems **35**, 30318–30332 (2022)
30. Abdin, M., Jacobs, S.A., Awan, A.A., Aneja, J., Awadallah, A., Awadalla, H., Bach, N., Bahree, A., Bakhtiari, A., Behl, H., et al.: Phi-3 technical report: A highly capable language model locally on your phone. arXiv preprint arXiv:2404.14219 (2024)
31. Timiryasov, I., Tastet, J.-L.: Baby llama: knowledge distillation from an ensemble of teachers trained on a small dataset with no performance penalty. arXiv preprint arXiv:2308.02019 (2023)
32. Dai, C., Li, K., Zhou, W., Hu, S.: Beyond Imitation: Learning Key Reasoning Steps from Dual Chain-of-Thoughts in Reasoning Distillation (2024). https://arxiv.org/abs/2405.19737
33. Luo, W., Fan, R., Li, Z., Du, D., Wang, Q., Chu, X.: Benchmarking and dissecting the nvidia hopper gpu architecture (2024). URL https://arxiv. org/abs/2402.13499
34. Rajpurkar, P., Zhang, J., Liu, K., Liang, P.: SQuAD: 100,000+ Questions for Machine Comprehension of Text (2016). https://arxiv.org/abs/1606.05250
35. Parrish, A., Chen, A., Nangia, N., Padmakumar, V., Phang, J., Thompson, J., Htut, P.M., Bowman, S.R.: BBQ: A Hand-Built Bias Benchmark for Question Answering (2022). https://arxiv.org/abs/2110.08193
36. Gou, Y., Liu, Z., Chen, K., Hong, L., Xu, H., Li, A., Yeung, D.-Y., Kwok, J.T., Zhang, Y.: Mixture of cluster-conditional lora experts for vision-language instruction tuning. arXiv preprint arXiv:2312.12379 (2023)
37. Deepmind, G.: Project Astra A universal AI agent that is helpful in everyday life (2024). https://deepmind.google/technologies/gemini/project-astra/
38. Sarlin, P.-E., DeTone, D., Malisiewicz, T., Rabinovich, A.: Superglue: Learning feature matching with graph neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4938–4947 (2020)
39. Laurençon, H., Tronchon, L., Cord, M., Sanh, V.: What matters when building vision-language models? arXiv preprint arXiv:2405.02246 (2024)

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 January 2025 doi:10.20944/preprints202501.0502.v1

40. Han, J., Du, L., Du, H., Zhou, X., Wu, Y., Zheng, W., Han, D.: Slim: Let llm learn more and forget less with soft lora and identity mixture. arXiv preprint arXiv:2410.07739 (2024)
41. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R.: Fairness Through Awareness (2011). https://arxiv.org/abs/1104.3913
42. Zhang, B., Jin, J., Fang, C., Wang, L.: Improved analysis of clipping algorithms for non-convex optimization. Advances in Neural Information Processing Systems **33**, 15511–15521 (2020)
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. Reid, M., Marrese-Taylor, E., Matsuo, Y.: Subformer: Exploring weight sharing for parameter efficiency in generative transformers. arXiv preprint arXiv:2101.00234 (2021)
45. Sanh, V.: Distilbert, a distilled version of bert: Smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108 (2019)
46. Faysse, M., Sibille, H., Wu, T., Viaud, G., Hudelot, C., Colombo, P.: Colpali: Efficient document retrieval with vision language models. arXiv preprint arXiv:2407.01449 (2024)
47. 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)
48. Qualcomm: Snapdragon 8 Gen 3 Mobile Platform. https://www.qualcomm.com/products/mobile/snapdragon/smartphones/snapdragon-8-series-mobile-platforms/snapdragon-8-gen-3-mobile-platform (2023)
49. Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J., (2011). https://doi.org/10.1561/2200000016
50. Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., Casas, D.d.l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al.: Mistral 7b. arXiv preprint arXiv:2310.06825 (2023)
51. BehnamGhader, P., Adlakha, V., Mosbach, M., Bahdanau, D., Chapados, N., Reddy, S.: Llm2vec: Large language models are secretly powerful text encoders. arXiv preprint arXiv:2404.05961 (2024)
52. Liu, Q., Wu, X., Zhao, X., Zhu, Y., Xu, D., Tian, F., Zheng, Y.: Moelora: An moe-based parameter efficient fine-tuning method for multi-task medical applications. arXiv preprint arXiv:2310.18339 (2023)
53. Adelani, D.I., Abbott, J., Neubig, G., D'souza, D., Kreutzer, J., Lignos, C., Palen-Michel, C., Buzaaba, H., Rijhwani, S., Ruder, S., *et al.*: Masakhaner: Named entity recognition for african languages. Transactions of the Association for Computational Linguistics **9**, 1116–1131 (2021)
54. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning, pp. 6105–6114 (2019). PMLR
55. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I.: Improving language understanding by generative pre-training. OpenAI blog (2018)
56. Beyer, L., Steiner, A., Pinto, A.S., Kolesnikov, A., Wang, X., Salz, D., Neumann, M., Alabdulmohsin, I., Tschannen, M., Bugliarello, E., et al.: Paligemma: A versatile 3b vlm for transfer. arXiv preprint arXiv:2407.07726 (2024)
57. Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A., Choi, Y.: Hellaswag: Can a machine really finish your sentence? arXiv preprint arXiv:1905.07830 (2019)
58. Research, A.M.L.: Introducing Apple's On-Device and Server Foundation Models. (2024). https://machinelearning.apple.com/research/introducing-apple-foundation-models Accessed October 2024
59. Team, G., Mesnard, T., Hardin, C., Dadashi, R., Bhupatiraju, S., Pathak, S., Sifre, L., Rivière, M., Kale, M.S., Love, J., et al.: Gemma: Open models based on gemini research and technology. arXiv preprint arXiv:2403.08295 (2024)
60. Mihaylov, T., Clark, P., Khot, T., Sabharwal, A.: Can a suit of armor conduct electricity? a new dataset for open book question answering. arXiv preprint arXiv:1809.02789 (2018)
61. Michel, P., Levy, O., Neubig, G.: Are sixteen heads really better than one? Advances in neural information processing systems **32** (2019)
62. Muralidharan, S., Sreenivas, S.T., Joshi, R., Chochowski, M., Patwary, M., Shoeybi, M., Catanzaro, B., Kautz, J., Molchanov, P.: Compact Language Models via Pruning and Knowledge Distillation (2024). https://arxiv.org/abs/2407.14679
63. Frantar, E., Alistarh, D.: Sparsegpt: Massive language models can be accurately pruned in one-shot. In: International Conference on Machine Learning, pp. 10323–10337 (2023). PMLR

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 January 2025 doi:10.20944/preprints202501.0502.v1

64. Ding, N., Chen, Y., Xu, B., Qin, Y., Zheng, Z., Hu, S., Liu, Z., Sun, M., Zhou, B.: Enhancing chat language models by scaling high-quality instructional conversations. arXiv preprint arXiv:2305.14233 (2023)
65. Nadeem, M., Bethke, A., Reddy, S.: StereoSet: Measuring stereotypical bias in pretrained language models. In: Zong, C., Xia, F., Li, W., Navigli, R. (eds.) 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. 5356–5371. Association for Computational Linguistics, Online (2021). https://doi.org/10.18653/v1/2021.acl-long.416 . https://aclanthology.org/2021.acl-long.416

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.