A Survey of Large and Small Language Models

Bojana Velichkovska, Jasmina Angelevska Kostadinoska, Dushko Stavrov, and Goran Jakimovski
Faculty of Electrical Engineering and Information Technologies, Ss. Cyril and Methodius University, Rugjer Boshkovikj 18, 1000 Skopje, North Macedonia
{bojanav, jasminaa, dushko.stavrov, goranj}@feit.ukim.edu.mk

Keywords: Artificial Intelligence, Natural Language Processing, Large Language Models, Small Language Models.

Abstract: The design and deployment of language models (LMs) is a significant advancement in the field of natural language processing (NLP). Their strong performance in understanding, interpreting, and generating human language has enabled their applicability in different fields, among which are some extremely sensitive areas as healthcare assistants, drug discovery and medical research, and education. With their potential impact, it is important to understand the way LMs operate and ensure their quality in undertaking these challenges. As such, two different approaches have emerged with specific applicability, namely large (LLMs) and small language models (SLMs). While LLMs have demonstrated high accuracy on generalized knowledge, SLMs have evolved to offer specialized insight for specific domains. Both LMs offer different benefits to their respective users. This study defines and examines the applicability of LLMs and SLMs, names the foundational research behind both, and presents a comparative analysis of the key architectural and methodological innovations and characteristics that have influenced their development, with focus on the newest and most used LMs.

1 INTRODUCTION

Research into natural language processing (NLP) has aimed to enable computers to comprehend, interpret and mimic human languages since its early development which incorporated machine translation and rule-based systems crafted by linguists to more data-driven approach through statistical methods and machine learning based on parsing and part-of speech tagging [1]. From there, the rise of deep learning and methods like recurrent neural networks (RNNs) and long short-term memory (LSTM) allowed computers to analyze human language without hands-on assistance [2]. However, as linguistically more complex challenges emerged so did major drawbacks, i.e., RNNs struggled with remembering information from earlier steps as sequences' size increased, and though LSTM offered better performance compared to RNNs, very long-range dependencies represent-ed a challenge as well. Moreover, the sequential nature of their architecture limited parallelization and had negative impact on training longer sequences.

The existing limitations were improved upon with the introduction of the transformer architecture [3]. Using a self-attention mechanism, transformers allowed the capture of long-term dependencies and were therefore more effective in handling long sequences of text. Since the self-attention approach processes an entire sequence at once, transformers can process data in parallel which offers scalability and speed when working with large datasets, and is computationally more efficient compared to previous models.

This established a stable foundation for the development of language models (LMs). Depending on the scale of the LMs, i.e., the number of trained parameters, there exist large LMs (LLMs) and small LMs (SLMs). With LMs' growing influence in applications from different domains, it is important to distinguish the applicability of LLMs and SLMs, as well as understand their scalability and efficiency.

This research focuses on analyzing and comparing the LLMs and SLMs. The rest of the paper is organized as follows. Section 2 presents the most known LLMs and their key characteristics. Section 3 lists the prevailing SLMs and their characteristics. Section 4 discusses and compares LLMs and SLMs, whereas Section 5 concludes the paper.

2 PREVAILING LLMS

LLMs build upon the transformer architecture to weigh the importance of different words in a sequence as means of understanding context. They leverage self-attention mechanisms to model long-range text dependencies through hundreds of billions of parameters trained on large datasets in order to obtain stronger language comprehension and generation performance [4]. Moreover, LLMs show emergent abilities which include following instructions for new tasks with or without being show examples of said task and solving complex tasks by dividing them into manageable subtasks [4].

The first impact of LLMs was the BERT (Bidirectional Encoder Representations from Transformers) model [5], which introduced masked language model-ling (with 15% of tokens masked) and bidirectional context learning. Its architecture resulted in improved performance in tasks like text classification and question answering. Subsequently, the GPT (Generative Pre-Training) series pioneered training models to predict the next token in a sequence [6]. The further development of the GPT family introduced GPT-3, with 175 billion parameters [7]. The most significant update in the model to its predecessor GPT-2 is the increase in the parameter number and the quantity of training data, which showcased that increasing model size could lead to significant leap in performance. The model redefined the field of NLP with its few-shot capabilities, i.e., the model could perform unseen tasks by having only few examples given as information, without explicit retraining. GPT-4 [8] enhanced safety measures compared to its predecessor and increased the context window size, whereas GPT-5 [9] increased user experience. PaLM [10] aimed to forecast sequences of actions over an extended period, and impacted translation tasks since the training process included 122 languages. LLaMA [11] introduced different layer normalization, new activation function, and instead of absolute positional embeddings employed rotary positional embeddings. The most significant changes in the LLaMA 2 and LLa-MA 3 models compared to their base model arise from the training dataset and the tokenization, showcasing that careful data curation and efficient tokenization can yield competitive performance even with fewer parameters than earlier models. In 2023, Google released Gemini [12] and boasted robust multimodal functionality and significant textual capabilities. DeepSeek [13] introduced lower computational costs, in spite of its size, as the model uses only a fraction of parameters in each interaction; whereas, Claude 4 [14] performed in-depth safety evaluation.

The characteristics of the most impactful LLMs, as described, are presented in Table 1, with model name, its release year, model size and its availability, as well as the novelty or impact of the model.

3 PREVAILING SLMS

Motivated by the need for efficiency, and accessibility in environments where power and latency make LLMs impractical to use, SLMs were introduced to retain the linguistic and reasoning abilities of LLMs, whilst reducing the computational cost and memory footprint. To achieve competitive performance with fewer resources, SLMs employ different optimization strategies, i.e., model architectures and efficient self-attention approximations, efficient pre-training and fine-tuning, and compression techniques (e.g., pruning, quantization, knowledge distillation) to reduce model size and latency without sacrificing accuracy [15].

Table 1: Overview of LLMs and key characteristics.

Model	Released	Size	Open Source	Key Characteristics
BERT	2018	340M	yes	Masked language modelling, and bidirectional context learning
GPT-3	2020	175B	no	Remarkable few-shot capabilities
PaLM	2022	540B	no	Reasoning and multilingual capabilities
LLaMA-1	2023	65B	yes	Rotary positional embeddings
LLaMA-2	2023	70B	yes	Highlights data importance
LLaMA-3	2024	405B	yes	Highlights data importance
Gemini	2023	3.25B	no	Multimodality
GPT-4	2023	1.76T (unofficially)	no	Larger context window, enhanced safety measures compared to GPT-3
DeepSeek	2024	671B	yes	Lower computational cost
GPT-5	2025	No data	no	User Experience improvement
Claude 4	2025	300-500B	no	Advanced safety evaluation

Table 2: Overview of SLMs and key characteristics.

Model	Released	Size	Open Source	Key Characteristics
DistilBERT	2019	66M	yes	Distilled BERT retaining ~95% performance
ALBERT	2019	12M–235M	yes	Parameter sharing, reduced embedding size
TinyBERT	2020	14.5M	yes	Two-stage distillation (general + task-specific)
MobileBERT	2020	25M	yes	Bottleneck structure for mobile efficiency
MiniLM	2020	22M	yes	Self-attention distillation
OPT-IML	2022	1.3B–30B	yes	Instruction-tuned compact transformer
MPT-7B	2023	7B	yes	Efficient pretraining, open-source release
Phi-2	2023	2.7B	yes	Textbook-quality data training
Gemma	2024	2B–7B	yes	Efficient scaling laws, lightweight deployment
SmolLM	2024	1.7B	yes	Optimized for edge devices
Mistral	2024	7B	yes	Sliding window attention, grouped-query attention

Initial breakthroughs in compact transformer architectures introduced Distil-BERT [16], a distilled version of BERT that reduced parameters by 40% whilst maintaining over 95% of its original performance. Next, ALBERT (A Lite BERT) [17] introduced cross-layer parameter sharing and factorised embeddings to achieve significant memory use reduction without major performance degradation. TinyBERT [18] optimised BERT for mobile and edge devices through two-stage distillation, i.e., general and task-specific distillation. MobileBERT [19] introduced a bottleneck structure and inverted-feed-forward layers, whereas MiniLM [20] leveraged self-attention distillation, leading to faster inference speeds and excellent performance on standard NLP benchmarks. The authors of [21] introduce OPT-IML (Optimized Pretrained Transformer for Instruction-based Machine Learning) series which showcase instruction tuning in compact architectures to retain reasoning ability in smaller models. MosaicML released MPT-7B in 2023 [22]. Their approach highlighted that careful dataset curation and efficient pre-training can produce competitive SLMs. The release of the Phi-2 model showed that fine-tuning with quality data can scale model performance [23]. Similarly, Gemma is trained on a limited dataset and obtains strong performance for language understanding and reasoning [24]. In 2024 and 2025, SmolLM2 reaffirmed the need for careful dataset curation and multistage training [25], whilst Mistral combined grouped-query attention for faster inference and sliding window attention to handle sequences of arbitrary length with lower computational costs [26].

The characteristics of the most impactful SLMs are presented in Table 2 with model name, its release year, model size and its availability, as well as the novelty or impact of the model.

4 LLMS AND SLMS DISCUSSION

LLMs have transformed how machines analyze and interpret human language, enabling multidisciplinary application with excellent performance in a variety of tasks (e.g., text generation, reasoning, translation, summarization, question answering, etc.). However, with their main goal being emulating human intelligence on a wider level, LLMs train a large parameter count invoking enormous computational requirements [27]. Moreover, a significant portion of LLMs can produce seemingly valid responses which upon analysis are found non-factual so-called hallucinations [28], and the scale and the data sources often used to train LLMs create another impactful problem, i.e., biased responses [29]. While advances have been made in ensuring data quality, the sheer volume of data that LLMs train on limits the improvements which can be made for both issues. From here, a parallel line of research focused on SLMs in order to achieve architectural compactness, computational efficiency, and low latency to counter the high computational demands for LLMs, as well as domain specialization to address hallucinations and biases. In this is one of the key operational differences between LLMs and SLMs, i.e., LLMs operate as resource-intensive generalizing apparatus and SLMs are their computationally-efficient and specialized counterparts designed for application in a constrained environment.

Table 3 shows a comparative analysis of LLMs and SLMs, providing additional differences between both on model scale and parameter count, model architecture, deployment, and performance context. LLMs employ deep and wide transformer stacks with extensive attention heads and context windows supported by specialized hardware and distributed training. In contrast, SLMs optimize for parameter

Table 3: Comparative analysis of LLMs and SLMs.

Characteristic	LLMs	SLMs
Scale and Parameters	Billions–trillions of parameters; massive training data	Millions–billions of parameters; optimised for compactness
Architecture	Deep transformer stacks; large attention windows	Parameter sharing, pruning, quantization, lightweight layers
Training Objective	Broad generalisation across diverse tasks	Domain specialisation and efficiency optimisation
Key Techniques	RLHF, instruction tuning, multimodal pretraining	Distillation, LoRA, quantisation, fine-tuning
Performance Focus	High accuracy, reasoning, and knowledge coverage	Low latency, interpretability, and adaptability
Deployment Context	Cloud and data centre environments	Edge, mobile, and domain-specific applications
Research Trend	Scaling laws, alignment, and multimodality	Efficiency scaling, hybrid integration, sustainability

reuse, pruning, and lightweight adaptation layers, ensuring scalability without overwhelming the computational demands. Methodologically, LLMs have led advancements in pretraining strategies and alignment techniques, whilst SLMs focus on knowledge transfer and compression to inherit capabilities from larger models and maintain agility. Research shows that LLMs generally outperform smaller models in reasoning, creativity, and general-domain understanding due to their extensive representational capacity. However, SLMs often achieve almost equal performance in domain-specific tasks particularly when fine-tuned with high-quality data. Recent trends indicate a convergence between large and small model philosophies. Techniques such as mixture-of-experts (MoE) architectures, modular training, and adaptive inference pipelines blur the boundary between LLMs and SLMs by combining scale with selectivity. The comparison highlights that the innovations in both LLMs and SLMs are mutually reinforcing rather than competitive. While LLMs continue to expand the boundaries of general intelligence through scale and data diversity, SLMs ensure that such capabilities become practically accessible and sustainable.

5 CONCLUSIONS

This survey presents and analyses three aspects of advancements in NLP. Firstly, we clarified the differences between LLMs and SLMs, in terms of scale, architecture, and performance. This distinction establishes a concept for understanding how scale influences capability, efficiency, and deployment considerations. Second, we reviewed most prevailing representatives of LLMs and SLMs in recent years, and listed the novelties they introduced. By surveying these developments, we provide insight into how both model families have evolved and how they continue to reshape research and practical applications. Finally, we compared and contrasted LLMs and SLMs across dimensions such as performance, resource requirements, and current research trends. Furthermore, we briefly touched upon issues in performance regarding hallucinations and biases. Given the rapid and ongoing pace of language model development, this survey aims to be an accessible and informative starting point for researchers entering the field.

ACKNOWLEDGMENTS

This research was funded by Ss. Cyril and Methodius University in Skopje (Project: Enabling Small Language Models for Efficient and Effective Application Using Retrieval-Augmented Generation, Number: 02-1298/9).

REFERENCES

[1] P. Johri, S. K. Khatri, A. T. Al-Taani, M. Sabharwal, S. Suvanov, and A. Kumar, “Natural Language Processing: History, Evolution, Application, and Future Work,” Lecture Notes in Networks and Systems, pp. 365–375, 2021, doi: 10.1007/978-981-15-9712-1_31.

[2] B. Ghojogh and A. Ghodsi, “Recurrent Neural Networks and Long Short-Term Memory Networks: Tutorial and Survey,” arXiv, Apr. 22, 2023. Available: https://arxiv.org/abs/2304.11461.

[3] A. Vaswani et al., “Attention Is All You Need,” arXiv, Jun. 12, 2017. Available: https://arxiv.org/abs/1706.03762.

[4] S. Minaee et al., "Large Language Models: A Survey," arXiv, Feb. 2024, doi: 10.48550/arxiv.2402.06196.

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," arXiv, Oct. 11, 2018. Available: https://arxiv.org/abs/1810.04805.

[6] R. He, A. Ravula, B. Kanagal, J. Ainslie, and G. Research, "RealFormer: Transformer Likes Residual Attention," Available: https://arxiv.org/pdf/2012.11747.

[7] T. B. Brown et al., "Language Models Are Few-Shot Learners," arXiv, vol. 4, no. 33, May 2020. Available: https://arxiv.org/abs/2005.14165.

[8] OpenAI, "GPT-4 Technical Report," arXiv:2303.08774, Mar. 2023, doi: 10.48550/arXiv.2303.08774.

[9] S. Wang, M. Hu, Q. Li, M. Safari, and X. Yang, "Capabilities of GPT-5 on Multimodal Medical Reasoning," arXiv, 2025. Available: https://arxiv.org/abs/2508.08224.

[10] A. Chowdhery et al., "PaLM: Scaling Language Modeling with Pathways," arXiv:2204.02311, Apr. 2022. Available: https://arxiv.org/abs/2204.02311.

[11] H. Touvron et al., "LLaMA: Open and Efficient Foundation Language Models," arXiv:2302.13971, Feb. 2023.

[12] R. Anil et al., "Gemini: A Family of Highly Capable Multimodal Models," arXiv, Dec. 18, 2023. Available: https://arxiv.org/abs/2312.11805.

[13] DeepSeek-AI et al., "DeepSeek-V3 Technical Report," arXiv, 2024. Available: https://arxiv.org/abs/2412.19437.

[14] Anthropic, "System Card: Claude Opus 4 & Claude Sonnet 4," 2025. Available: https://www-cdn.anthropic.com/6d8a8055020700718b0c49369f60816ba2a7c285.pdf.

[15] V. Nguyen et al., "A Survey of Small Language Models," arXiv, 2024. Available: https://arxiv.org/abs/2410.20011.

[16] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," arXiv, 2019. Available: https://arxiv.org/abs/1910.01108.

[17] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "ALBERT: A Lite BERT for Self-supervised Learning of Language Representations," arXiv:1909.11942, Feb. 2020. Available: https://arxiv.org/abs/1909.11942.

[18] X. Jiao et al., "TinyBERT: Distilling BERT for Natural Language Understanding," arXiv:1909.10351, Oct. 2020. Available: https://arxiv.org/abs/1909.10351.

[19] Z. Sun, H. Yu, X. Song, R. Liu, Y. Yang, and D. Zhou, "MobileBERT: a Compact Task-Agnostic BERT for Resource-Limited Devices," arXiv:2004.02984, Apr. 2020. Available: https://arxiv.org/abs/2004.02984.

[20] W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, and M. Zhou, "MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers," arXiv:2002.10957, Apr. 2020. Available: https://arxiv.org/abs/2002.10957.

[21] S. Iyer et al., "OPT-IML: Scaling Language Model Instruction Meta Learning through the Lens of Generalization," arXiv, Dec. 2022, doi: 10.48550/arxiv.2212.12017.

[22] "Introducing MPT-7B: A New Standard for Open-Source, Commercially Usable LLMs," Databricks, May 05, 2023. Available: https://www.databricks.com/blog/mpt-7b.

[23] A. Hughes, "Phi-2: The surprising power of small language models," Microsoft Research, Dec. 12, 2023. Available: https://www.microsoft.com/en-us/research/blog/phi-2-the-surprising-power-of-small-language-models/.

[24] Gemma Team et al., "Gemma: Open Models Based on Gemini Research and Technology," arXiv, Mar. 13, 2024. Available: https://arxiv.org/abs/2403.08295.

[25] A. L. Ben et al., "SmolLM2: When Smol Goes Big – Data-Centric Training of a Small Language Model," arXiv, 2025. Available: https://arxiv.org/abs/2502.02737.

[26] A. Q. Jiang et al., "Mistral 7B," arXiv, Oct. 10, 2023. Available: https://arxiv.org/abs/2310.06825.

[27] J. Camilo, "Efficient Strategy for Improving Large Language Model (LLM) Capabilities," arXiv, 2025. Available: https://www.arxiv.org/abs/2508.04073.

[28] L. Huang et al., "A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions," arXiv, Nov. 2023, doi: 10.48550/arxiv.2311.05232.

[29] R. Ranjan, S. Gupta, and S. N. Singh, "A Comprehensive Survey of Bias in LLMs: Current Landscape and Future Directions," arXiv, 2024. Available: https://arxiv.org/abs/2409.16430.