iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH
IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 Issue: IX Month of publication: September 2025

DOI: https://doi.org/10.22214/ijraset.2025.73913

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue IX Sep 2025- Available at www.ijraset.com

Redefining Natural Language Processing Efficiency: The Rise of Small Language Models

Dr. Goldi Soni Mam[1], Rohan Sahu[2], Bhavya Bisen[3]
[1]*Assistant Professor, Amity University Chhattisgarh*
[2, 3]*Student, BTech CSE, Amity University Chhattisgarh*

Abstract: The rapid advancement of Natural Language Processing (NLP) has been driven by large language models (LLMs), but their extensive computational and memory requirements pose significant challenges. Small Language Models (SLMs) are emerging as an efficient alternative, offering competitive performance with reduced resource demands. This paper explores the architecture, training techniques, and optimization strategies that enable SLMs to achieve remarkable efficiency. It reviews key breakthroughs, including knowledge distillation, parameter pruning, and quantization, which contribute to their lightweight design. Additionally, the paper highlights practical applications where SLMs outperform larger models in terms of speed, adaptability, and deployment feasibility, particularly in resource-constrained environments. The analysis aims to present SLMs as a promising direction for sustainable, accessible, and effective NLP solutions.

Keywords: Language Processing , Training techniques , Pruning , Lightweight design , Resource-constrained.

I. INTRODUCTION

In recent years, the field of Natural Language Processing (NLP) has seen significant advancements with the emergence of Large Language Models (LLMs). While LLMs have demonstrated remarkable capabilities in text generation, comprehension, and reasoning, their high computational requirements, energy consumption, and resource-intensive nature have raised concerns regarding efficiency and accessibility. These challenges have led to an increasing interest in Small Language Models (SLMs)—compact yet powerful models designed to deliver efficient NLP performance with lower computational costs.

This research paper presents a comprehensive review of various studies on SLMs and their role in enabling efficient NLP. By analyzing existing research, we explore the trade-offs between model size, accuracy, efficiency, and applicability in real-world scenarios. We discuss innovations in model architectures, training techniques, and optimization strategies that contribute to the effectiveness of SLMs while maintaining competitive performance compared to larger counterparts.

A. Background and Motivation

Overview of LLMs and Their Limitations

Large Language Models (LLMs), such as GPT-4, PaLM, and LLaMA, have revolutionized NLP by achieving state-of-the-art performance across various tasks. These models leverage billions of parameters, enabling them to understand complex linguistic patterns and generate human-like responses. However, their widespread adoption is hindered by several challenges:

- High Computational Costs: Training and deploying LLMs require significant computing resources, making them inaccessible to many researchers and developers.
- Energy Consumption: Running LLMs demands substantial energy, raising concerns about sustainability and environmental impact.
- Latency and Storage Constraints: LLMs require large-scale infrastructure, limiting their usability in real-time applications and edge devices.

B. The Need for Efficiency in NLP

As AI adoption expands across industries, the demand for **efficient NLP models** has increased. Organizations seek solutions that provide high-quality language processing while minimizing costs and environmental impact. Small Language Models (SLMs) offer a promising alternative by reducing computational requirements while maintaining strong performance, making NLP more accessible and scalable.

C. Defining Small Language Models (SLMs)

1) Characteristics of SLMs

SLMs are designed to offer a balance between model size, computational efficiency, and performance. They typically possess the following characteristics:

- Lower Parameter Count: SLMs contain significantly fewer parameters than LLMs, often in the range of millions instead of billions.
- Optimized Architectures: These models leverage pruning, quantization, knowledge distillation, and efficient transformer variants to maintain performance with reduced complexity.
- Faster Inference Times: Due to their compact size, SLMs are well-suited for real-time applications and deployment on edge devices.

2) Comparison with LLMs

While LLMs excel in capturing deep contextual relationships, SLMs prioritize efficiency without compromising essential NLP capabilities. Studies indicate that task-specific fine-tuned SLMs can achieve comparable performance to LLMs in various NLP tasks while significantly reducing resource requirements.

II. LITERATURE REVIEW

The research landscape of Small Language Models (SLMs) and Efficient NLP builds upon foundational works in deep learning, Transformer architectures, and model optimization techniques. This literature review categorizes the key contributions of existing studies into four main areas: Transformer architectures and efficiency improvements, model compression techniques, quantization and pruning methods, and real-world applications of SLMs.

(Sutherland et al., 2022) The paper by Blagec et al. provides a comprehensive analysis of performance metrics used in natural language processing (NLP). Traditional metrics like BLEU and ROUGE, despite their widespread use, have been criticized for low correlation with human judgment and poor transferability across tasks. The authors curated over 3500 performance results from 'Papers with Code' to analyze the prevalence and effectiveness of various metrics. They found that most metrics used in NLP may inadequately reflect model performance and highlighted inconsistencies in metric reporting. The study emphasizes the need for clear and unambiguous reporting of metrics to improve transparency and reproducibility. Despite the introduction of superior metrics like BERTScore and METEOR, traditional metrics remain dominant. The authors recommend using multiple metrics and ensuring detailed reporting to enhance the informative value of performance results. They also suggest the development of a taxonomic hierarchy of performance metrics to systematize definitions and applications. The paper concludes that large comparative evaluation studies are necessary to better understand the suitability of different metrics for various NLP tasks.

(Bommasani et al., 2021) The report explores the paradigm shift in AI with the rise of foundation models like BERT, GPT-3, and DALL-E, which are trained on broad data using self-supervision and can be adapted to various downstream tasks. These models exhibit new emergent capabilities due to their scale, leading to both opportunities and risks. The report discusses the technical principles, applications, and societal impacts of foundation models, emphasizing the need for interdisciplinary collaboration. The historical context of AI development is traced from machine learning to deep learning and now to foundation models. The report addresses the social impact of foundation models, the importance of data curation, and the need for responsible deployment and monitoring. Foundation models have revolutionized natural language processing (NLP) by shifting the focus from bespoke architectures for different tasks to leveraging these models for various applications. These models have shown significant success in language generation tasks such as summarization and dialogue generation. The adaptability of foundation models to handle language variation and multilinguality remains an open question. Foundation model-powered applications must balance user agency and values with the benefits of AI automation. The philosophy of understanding in foundation models involves debates on whether these models can truly understand language.Foundation models have the potential to transform healthcare, law, and education by improving efficiency, accuracy, and accessibility. The text discusses the challenges and opportunities of using foundation models in education. The text discusses the challenges and opportunities in adapting foundation models for various tasks, emphasizing the need for continual learning. Foundation models face significant challenges related to demographic biases and robustness. The text discusses the biases and risks associated with foundation models, emphasizing the need for new protocols to address data and model biases.

(Fedus et al., 2021)The Switch Transformer is a sparsely-activated model designed to address the complexity, communication costs, and training instability of Mixture of Experts (MoE) models.

It simplifies the MoE routing algorithm, reduces communication and computational costs, and introduces training techniques to mitigate instabilities. The model achieves significant improvements in pre-training speed and efficiency, particularly in multilingual settings, and scales up to trillion parameter models. Key contributions include the Switch Transformer architecture, improved training and fine-tuning techniques, and successful distillation of large sparse models into smaller dense models. The model demonstrates superior scaling properties, achieving better performance with increased parameters while maintaining constant computational costs. It also shows significant improvements in downstream tasks, including fine-tuning and multilingual learning, and offers effective strategies for combining data, model, and expert-parallelism. The text discusses various strategies for partitioning model weights and data across multiple cores in large-scale models, focusing on Switch Transformers. It explains the use of model and data parallelism, expert and data parallelism, and the combination of expert, model, and data parallelism. The partitioning strategies involve splitting model weights and data tensors across cores to balance computation and memory usage. The text also highlights the design and performance of large Switch Transformer models, including those with up to 1.6 trillion parameters. These models show improvements in pre-training and fine-tuning tasks, although training stability remains a challenge. The text compares the performance of Switch Transformers with T5 models, noting that Switch Transformers are more sample efficient and faster.

(Hoffmann et al., 2022)Hoffmann et al. (2022) challenge the prevailing trend of scaling language models by increasing parameter size while keeping training data constant. They propose that for compute-optimal training, model size and training tokens should scale equally, contradicting earlier findings by Kaplan et al. (2020). Their study, based on training over 400 models, leads to the creation of Chinchilla, a 70B-parameter model trained on 1.4T tokens, which outperforms larger models like Gopher (280B), GPT-3 (175B), and Megatron-Turing NLG (530B) on multiple benchmarks. Chinchilla achieves 67.5% accuracy on MMLU, exceeding Gopher’s performance by over 7%. This approach not only improves efficiency but also reduces fine-tuning and inference costs, making models more practical for real-world applications. The findings emphasize the importance of high-quality, large-scale datasets rather than just increasing model size. This shift in strategy has significant implications for AI sustainability, lowering computational and energy costs while maintaining superior performance. The study highlights a crucial paradigm shift in AI, demonstrating that smaller, well-trained models can outperform larger, undertrained ones.

(Sanh et al., 2019)The paper proposes a Patient Knowledge Distillation (Patient-KD) approach to compress large pre-trained language models like BERT into lightweight models without sacrificing performance. Unlike traditional methods that use only the final layer's output, Patient-KD leverages multiple intermediate layers for incremental knowledge extraction. Two strategies, PKD-Last and PKD-Skip, are introduced to exploit the rich information in the teacher model's hidden layers. Experiments on various NLP tasks demonstrate that Patient-KD achieves superior performance and better generalization compared to standard distillation methods, with significant gains in training efficiency and storage reduction. The approach is validated on tasks such as Sentiment Classification, Paraphrase Similarity Matching, Natural Language Inference, and Machine Reading Comprehension, showing that it maintains comparable accuracy to the original large models.

(Tanaka et al., 2020)The paper discusses a novel approach to pruning neural networks without using any data, called Iterative Synaptic Flow Pruning (SynFlow). This method aims to identify highly sparse trainable subnetworks at initialization, avoiding the need for expensive training and pruning cycles. The authors introduce a conservation law to explain why existing gradient-based pruning algorithms suffer from layer-collapse, which renders a network untrainable.

SynFlow preserves the total flow of synaptic strengths through the network, achieving state-of-the-art results across various models and datasets. The algorithm is data-agnostic and consistently outperforms existing methods, challenging the paradigm that data is necessary for effective pruning at initialization. The study also highlights the importance of avoiding layer-collapse and demonstrates that SynFlow can reach maximal critical compression. The authors benchmark SynFlow against other pruning methods, showing its superior performance in high compression regimes. They also provide theoretical insights into why SynFlow avoids layer-collapse and propose future directions for improving pruning algorithms. The work has implications for increasing the energy efficiency of neural networks and facilitating their deployment on edge devices.

(Zafrir et al., 2019)The paper discusses the challenges of deploying large pre-trained Transformer-based language models like BERT in production environments due to their high computational and memory requirements. The authors propose a method for compressing BERT by applying quantization-aware training during the fine-tuning phase, which reduces the model size by 4× with minimal accuracy loss. This method involves quantizing all GEMM operations in BERT's Fully Connected and Embedding layers to 8-bit integers, achieving a significant reduction in memory footprint and potential acceleration in inference speed. The quantized model maintains 99% accuracy compared to the FP32 version across various NLP tasks. The authors implemented this method using symmetric linear quantization and fake quantization during training to simulate quantization errors. They evaluated the quantized BERT on the GLUE benchmark and SQuAD dataset, demonstrating minimal accuracy loss and significant efficiency gains. The paper highlights the potential for deploying efficient, low-latency NLP applications on various hardware platforms using this quantization method. Future work includes exploring additional model compression techniques to further enhance BERT's efficiency.

(Frankle & Carbin, 2019)The paper introduces the Lottery Ticket Hypothesis, suggesting that dense, randomly-initialized neural networks contain smaller subnetworks (winning tickets) that can be trained to achieve comparable test accuracy to the original network. The authors present an algorithm to identify these winning tickets and demonstrate that these subnetworks can be significantly smaller and still perform effectively. Experiments on fully-connected and convolutional networks for MNIST and CIFAR10 support the hypothesis, showing that winning tickets can learn faster and achieve higher test accuracy. The paper explores the implications for improving training performance, designing better networks, and understanding neural network optimization. It discusses the importance of initialization and structure in the success of winning tickets and proposes that overparameterized networks are easier to train due to more potential winning tickets. The document analyzes the performance of various neural network architectures under different pruning strategies and hyperparameters. It explores the impact of iterative pruning, learning rates, and initialization methods on test accuracy and early-stopping iterations. Key findings include that global pruning generally outperforms layer-wise pruning, especially for deeper networks.The document discusses the robustness of winning tickets to Gaussian noise and the connectivity patterns of pruned networks. Experiments reveal that lower learning rates and warmup strategies can help in finding smaller winning tickets that maintain high accuracy.

(Jiao et al., 2020)TinyBERT is a distilled version of BERT designed to reduce model size and inference time while maintaining accuracy. The novel Transformer distillation method transfers knowledge from a large BERT model to a smaller TinyBERT model. TinyBERT employs a two-stage learning framework, performing distillation at both pre-training and task-specific stages. TinyBERT4, with 4 layers, achieves 96.8% of BERTBASE's performance on the GLUE benchmark, while being 7.5 times smaller and 9.4 times faster. TinyBERT6, with 6 layers, performs on par with BERTBASE. The method includes attention-based and hidden states-based distillation, ensuring effective knowledge transfer. Data augmentation is used to enhance task-specific distillation. Experiments show that TinyBERT significantly outperforms other state-of-the-art baselines in BERT distillation. The approach is effective for various NLP tasks, providing a way to deploy BERT-based models on resource-limited devices.

(Yao et al., 2022)ZeroQuant is an efficient post-training quantization approach for compressing large Transformer-based models, featuring a fine-grained hardware-friendly quantization scheme for weights and activations. It includes a novel layer-by-layer knowledge distillation algorithm (LKD) that does not require original training data and an optimized quantization system backend to eliminate quantization/dequantization overhead. ZeroQuant can reduce precision to INT8 for weights and activations with minimal accuracy impact, achieving significant speedups and memory footprint reductions. It has been successfully applied to large models like GPT-J6B and GPT-NeoX20B, demonstrating substantial efficiency improvements. The methodology includes group-wise quantization for weights, token-wise quantization for activations, and optimized inference kernels. Empirical results show that ZeroQuant maintains accuracy while significantly reducing inference latency and memory requirements. The approach is scalable and effective even without access to original training data, using alternative datasets for knowledge distillation. Using Wikipedia data can further enhance accuracy and reduce perplexity (PPL), making it a viable alternative when the original dataset is unavailable.

(Lan et al., 2020)The paper presents ALBERT, a lighter version of BERT designed to address memory and training speed limitations. ALBERT uses two parameter-reduction techniques: factorized embedding parameterization and cross-layer parameter sharing, which significantly reduce the number of parameters without compromising performance. Additionally, ALBERT introduces a self-supervised loss for sentence-order prediction (SOP) to improve inter-sentence coherence modeling. These innovations allow ALBERT to achieve state-of-the-art results on benchmarks like GLUE, RACE, and SQuAD with fewer parameters than BERT-large. The paper also discusses the impact of network depth and width on performance, finding diminishing returns with increased layers and hidden sizes. The authors highlight the importance of parameter efficiency and propose future research directions to further enhance training and inference speed.

(Dai et al., 2020)The Funnel-Transformer is designed to improve the efficiency of language processing by reducing the redundancy in maintaining a full-length token-level presentation. It compresses the sequence of hidden states, reducing computation costs and allowing for deeper or wider models. The model can recover deep representations for each token via a decoder, making it suitable for tasks requiring token-level predictions. Empirical results show that Funnel-Transformer outperforms standard Transformers on various sequence-level tasks with comparable or fewer FLOPs. The architecture retains the Transformer’s modular structure but introduces pooling to reduce sequence length, enhancing efficiency without compromising performance. The model is particularly effective for tasks like text classification, language understanding, and reading comprehension. The Funnel-Transformer also shows promise in handling long texts and complex reasoning tasks, as demonstrated by its performance on the RACE dataset. However, it may not perform as well on tasks requiring detailed token-level information, such as SQuAD. Future research could focus on optimizing the compression scheme and block layout design to further enhance the model's efficiency and performance.

(Wang et al., 2020)The paper introduces Linformer, a novel transformer model that reduces the complexity of self-attention from O(n^2) to O(n) in both time and space. Traditional transformer models, while effective, are computationally expensive due to their self-attention mechanism. Linformer approximates self-attention using a low-rank matrix, significantly improving efficiency without sacrificing performance. The authors demonstrate that self-attention can be decomposed into multiple smaller attentions through linear projections, forming a low-rank factorization. Experiments show that Linformer performs comparably to standard transformers on various NLP tasks while offering substantial speed and memory advantages. The model's efficiency is further enhanced by techniques like parameter sharing and nonuniform projected dimensions. Linformer maintains performance even with longer sequences, making it suitable for real-world applications. Theoretical and empirical analyses support the model's linear-time complexity, and additional techniques can further optimize performance and efficiency. The study concludes that Linformer offers a practical solution to the computational challenges of transformer models.

(Sun et al., 2020)MobileBERT is a compact version of BERT designed for resource-limited devices, maintaining task-agnostic properties for various NLP tasks. It is 4.3 times smaller and 5.5 times faster than BERTBASE, achieving competitive results on benchmarks. MobileBERT uses bottleneck structures and balances self-attentions with feed-forward networks. Training involves a specially designed teacher model, IB-BERTLARGE, and knowledge transfer. MobileBERT achieves a GLUE score of 77.7 and performs well on SQuAD tasks. It incorporates operational optimizations like NoNorm and relu activation to reduce latency. Progressive knowledge transfer is used to train MobileBERT effectively. The model's architecture includes deep and thin layers with stacked feed-forward networks to maintain performance.

(Kitaev et al., 2020)The Reformer model, introduced in a conference paper at ICLR 2020, aims to improve the efficiency of Transformer models, which are widely used in natural language processing but are resource-intensive. The Reformer achieves this by using reversible residual layers and locality-sensitive hashing (LSH) attention, reducing memory complexity from O(L2) to O(L log L). Reversible layers allow storing activations only once, while LSH attention focuses on the nearest keys, making the model more memory-efficient and faster on long sequences. The model also introduces shared-QK attention, where queries and keys are identical, and chunking in feed-forward layers to further save memory. The Reformer is particularly effective for long sequences, making it suitable for tasks like text generation, time-series forecasting, and image and video generation. The paper also discusses the impact of various techniques on performance, showing that the Reformer can handle large models efficiently. Overall, the Reformer combines the modeling capacity of Transformers with improved efficiency, making it accessible for broader use in research and industry.

(Zhang et al., 2024)TinyLlama is a compact 1.1B parameter open-source language model trained on approximately 3 trillion tokens over multiple epochs. It builds on Llama 2's architecture and incorporates optimizations like FlashAttention and Lit-GPT for enhanced computational efficiency. Despite its smaller size, TinyLlama achieves strong performance in various NLP tasks, outperforming similarly sized open-source models such as OPT-1.3B and Pythia-1.4B.The model's pretraining data includes a mix of SlimPajama and StarCoder datasets, containing natural language and code data. TinyLlama employs a decoder-only Transformer architecture with features like Rotary Positional Embedding, SwiGLU activation, and Grouped-query Attention to optimize performance. To improve efficiency, the training process integrates Fully Sharded Data Parallel (FSDP) and FlashAttention-2, reducing GPU hours compared to similar models.

(Wolf et al., 2020)The Hugging Face Transformers Library is an open-source framework for developing and using Transformer-based NLP models. It provides a unified API for state-of-the-art models like BERT, GPT, and T5, making them accessible for research and industry. The library features a Model Hub for accessing and sharing pretrained models, reducing training efforts.It supports diverse NLP tasks such as text classification, translation, question answering, and summarization. With compatibility for PyTorch and TensorFlow, it ensures seamless interoperability. Optimized Rust-based tokenizers enable fast text processing.The library includes tools for training, fine-tuning, and deployment, with ONNX and CoreML support for edge-device inference. Its community-driven development ensures continuous enhancements, making it vital for advancing NLP research and applications in academia and industry.

(Microsoft, 2023)Phi-2 is a small language model (SLM) by Microsoft, proving that high performance is possible without massive scale. With 2.7 billion parameters, it rivals much larger models using high-quality training data and efficient scaling.Building on the Phi series, it excels in coding and natural language tasks. Its success comes from "textbook-quality" data instead of generic web sources, ensuring better learning. Techniques like weight reuse across different scales enhance efficiency.Microsoft refined training with synthetic data generation and filtering, boosting performance at lower costs. Phi-2 surpasses previous models in reasoning and coding while remaining compact. It marks a shift in AI, proving small, well-trained models can be powerful and cost-effective.

(Microsoft, 2023)Phi-2 is a small language model (SLM) by Microsoft, proving that high performance is possible without massive scale. With 2.7 billion parameters, it rivals much larger models using high-quality training data and efficient scaling.Building on the Phi series, it excels in coding and natural language tasks. Its success comes from "textbook-quality" data instead of generic web sources, ensuring better learning. Techniques like weight reuse across different scales enhance efficiency.Microsoft refined training with synthetic data generation and filtering, boosting performance at lower costs. Phi-2 surpasses previous models in reasoning and coding while remaining compact. It marks a shift in AI, proving small, well-trained models can be powerful and cost-effective.

(Jiang et al., 2023)Mistral 7B is a high-performance 7-billion-parameter language model optimized for efficiency and accuracy. It outperforms Llama 2 13B and Llama 1 34B in reasoning, math, and code generation while remaining compact.The model uses Grouped-Query Attention (GQA) for faster inference and Sliding Window Attention (SWA) to handle long sequences efficiently. It includes an instruction-tuned variant, Mistral 7B – Instruct, which surpasses Llama 2 13B in human and automated evaluations.Mistral 7B excels in benchmarks for commonsense reasoning, world knowledge , reading comprehension, and coding. It supports fine-tuning and deployment on AWS, GCP, and Azure.The model integrates system prompts and self-reflection for enhanced safety and content moderation. Released under the Apache 2.0 license, Mistral 7B sets a new standard for efficient and powerful AI models.

(Strubell et al., 2019)The survey on efficient NLP explores methods to enhance performance while minimizing resource consumption, including computation, memory, and energy. Large-scale models demand significant resources, making efficiency crucial for accessibility and sustainability.Efficiency improvements are categorized into data optimization (filtering, active learning, curriculum learning), model design (efficient attention, sparse modeling, parameter efficiency), and training techniques (pre-training and fine-tuning optimizations).Key inference efficiency techniques include pruning, quantization, and knowledge distillation. Hardware-specific optimizations and co-design strategies enhance model performance across computing platforms.The study also evaluates efficiency metrics like FLOPs, power consumption, and carbon footprint, emphasizing the need for a balance between efficiency and model effectiveness. It serves as a guide for researchers to develop practical, resource-efficient NLP systems.

(Zhang et al., 2022)TextPruner is an open-source toolkit for pruning pre-trained language models efficiently, reducing size and inference time without retraining. It supports structured pruning methods like vocabulary and transformer pruning.Vocabulary pruning removes rarely used tokens, while transformer pruning eliminates less important attention heads and feed-forward neurons. The toolkit offers both Python API and CLI for ease of use.TextPruner enables self-supervised pruning, eliminating the need for labeled data. It supports NLP tasks such as text classification and machine reading comprehension.By using optimization-free pruning techniques, TextPruner balances efficiency and performance. It enhances model compression and deployment, making deep learning more accessible.

(Fedus et al., 2021)Switch Transformers use a scalable, sparsely activated Mixture of Experts (MoE) architecture, activating only a subset of parameters per input while keeping computational costs constant.They simplify MoE routing, reducing communication overhead and improving training stability. Lower precision (bfloat16) training enhances efficiency.Pre-trained on the Colossal Clean Crawled Corpus (C4), they achieve up to 7x faster pre-training speeds than T5 models. They excel in multilingual tasks, improving performance across 101 languages.Switch Transformers outperform dense models in reasoning and knowledge-based tasks while allowing efficient distillation into smaller models. By combining expert, model, and data parallelism, they scale efficiently while maintaining computational feasibility.

(Sutherland et al., 2022)The paper examines efficiency metrics in NLP, highlighting the shortcomings of traditional metrics like BLEU and ROUGE. These metrics often fail to align with human judgment and lack transferability across tasks.By analyzing over 3,500 performance results from "Papers with Code," the study reveals inconsistencies in reporting that hinder model comparison and transparency. Alternative metrics like BERTScore and METEOR offer improvements but remain underutilized.The authors advocate for standardized metrics, clearer reporting, and more reliable evaluation methods. Future research should focus on adaptable benchmarking frameworks that consider efficiency, fairness, and real-world applications.

(Wu et al., 2020)FastFormers optimizes Transformer-based NLP models for efficient inference using knowledge distillation, structured pruning, and model quantization.Knowledge distillation compresses large models into smaller ones with minimal accuracy loss. Structured pruning removes less important attention heads and feed-forward layers. Model quantization speeds up execution using lower-precision arithmetic on CPUs and GPUs.FastFormers improves speed by up to 233x on CPUs and 12.4x on GPUs while reducing energy consumption. It enables cost-effective large-scale deployment, making Transformer models more efficient and accessible.

(Liu et al., 2021)EdgeBERT is a latency-aware optimization framework for efficient NLP inference on edge devices. It reduces computational costs while maintaining accuracy using early exit mechanisms, DVFS, and adaptive attention span techniques.Entropy-based early exit prediction allows sentences to exit inference once a confidence threshold is met, saving energy. Network pruning and floating-point quantization further improve efficiency.EdgeBERT integrates specialized hardware, including a fast-switching voltage regulator and non-volatile memory storage. It achieves up to 7x energy savings over conventional BERT inference, making NLP models more practical for real-time edge applications.

(Hoffmann et al., 2022)The study investigates the optimal balance between model size and training data for compute-efficient large language models. Current models are often undertrained due to a focus on scaling parameters while keeping training data constant.Through training over 400 models, the authors find that optimal scaling requires increasing both model size and training tokens equally. They validate this by training Chinchilla (70B parameters, 1.4T tokens), which outperforms larger models like Gopher (280B), GPT-3 (175B), and Megatron-Turing NLG (530B).Chinchilla achieves superior results in NLP benchmarks while requiring less compute for fine-tuning and inference. The study challenges previous scaling laws, advocating for smaller, well-trained models as more efficient alternatives for large-scale AI applications.

(Wang et al., 2020)MiniLM is a lightweight Transformer model that uses deep self-attention distillation to compress large pre-trained models like BERT while retaining high performance. Instead of layer-to-layer distillation, it mimics only the last Transformer layer's self-attention module, making student models more flexible.The approach transfers both self-attention distributions and value relations to improve knowledge retention. A teacher assistant model helps bridge the gap between large teacher models and small student models, enhancing the distillation process.MiniLM achieves up to 99% of BERT's accuracy on benchmarks like SQuAD 2.0 and GLUE while using 50% fewer parameters. It also supports multilingual tasks, showing competitive performance in cross-lingual understanding. The model is efficient, fast, and well-suited for real-world applications with limited computational resources.

(Wang et al., 2020)Hardware-Aware Transformers (HAT) is an optimization framework that designs efficient Transformer models tailored for specific hardware. It employs Neural Architecture Search (NAS) to create models optimized for CPUs, GPUs, and edge devices. HAT constructs a large design space with arbitrary encoder-decoder attention and heterogeneous layers. It trains a weight-sharing SuperTransformer, from which smaller SubTransformers are extracted. An evolutionary search with latency constraints finds the best model for each hardware platform.HAT achieves up to 3× speedup and 3.7× model compression without accuracy loss. It enables low-latency NLP deployment on resource-limited devices, making Transformer models more accessible and efficient.

(Gou et al., 2021)Knowledge distillation (KD) is a model compression technique where a smaller student model learns from a larger teacher model while maintaining high accuracy. It transfers knowledge using different approaches, such as response-based, feature-based, and relation-based distillation.KD improves model efficiency, making deep learning models more deployable on resource-constrained devices. It includes offline, online, and self-distillation methods, with various architectures like multi-teacher, adversarial, and cross-modal distillation.Applications of KD span across computer vision, speech recognition, and NLP. The survey also discusses challenges, including model capacity gaps, optimization biases, and dataset dependencies. Future research focuses on improving KD's effectiveness, automation, and generalization.

(Sun et al., 2019)Patient Knowledge Distillation (PKD) is a model compression technique designed to make BERT more efficient while maintaining high accuracy. Unlike traditional distillation, which learns only from the final layer, PKD extracts knowledge from multiple intermediate layers for better generalization.PKD follows two strategies: PKD-Last (learning from the last k layers) and PKD-Skip (learning from every k layers). These methods allow the student model to absorb richer information from the teacher model.PKD significantly improves training efficiency while maintaining strong performance on NLP tasks like sentiment analysis, paraphrase matching, and reading comprehension. It enables faster inference and storage efficiency, making BERT more practical for real-world applications.

III. COMPARISON OF PAST RESEARCH PAPERS

The comparison highlights BERT's strong performance but high resource use, while DistilBERT and TinyBERT achieve efficiency through distillation, Hoffmann et al. optimize scaling, and Hugging Face simplifies model deployment.

Table 1. Comparison of past Research Papers

Research Paper	Authors	Year	Objective	Outcome	Limitations	Future Scope
Training Compute-Optimal Large Models	Hoffmann et al.	2022	Optimize model size and data balance under a fixed compute budget	Smaller models trained on more data outperform larger ones trained on less data	Assumes ideal training conditions	Extend to diverse architectures and datasets
TinyBERT	Jiao et al.	2020	Compress BERT via knowledge distillation for faster inference	>96% of BERT's performance while being 7.5x smaller and 9x faster	Struggles with complex reasoning tasks	Improve performance on complex NLP tasks, expand to more domains
Hugging Face Transformers Library	Wolf et al.	2020	Provide an open-source library with state-of-the-art Transformer models	Simplifies integration and deployment of top NLP models like BERT, GPT-2, and RoBERTa	Large models still require significant resources	Expand model support, optimize for edge devices
BERT	Devlin et al.	2019	Pre-train deep bidirectional representations from unlabeled text	Achieved state-of-the-art performance on GLUE, SQuAD, and other NLP benchmarks	High resource consumption and slow inference	Improve efficiency and adaptability to new tasks

IV. CONCLUSION

The evolution of Natural Language Processing has reached a pivotal moment where efficiency is as critical as performance. While Large Language Models (LLMs) have set benchmarks in linguistic capabilities, their resource-intensive nature limits widespread adoption. Small Language Models (SLMs) emerge as a compelling solution, offering a balance between computational efficiency and task-specific accuracy.

This review highlights how innovations such as knowledge distillation, pruning, quantization, and optimized transformer architectures empower SLMs to rival their larger counterparts. From TinyBERT and MobileBERT to Phi-2 and TinyLlama, the landscape is rich with models that demonstrate high performance in constrained environments. These advancements not only democratize access to NLP technologies but also pave the way for sustainable AI development.

As industries increasingly demand scalable, low-latency, and energy-efficient solutions, SLMs stand out as the future of practical NLP. Continued research into training strategies, hardware-aware design, and evaluation metrics will further enhance their capabilities, making intelligent language systems more inclusive and environmentally responsible.

REFERENCES

[1] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. Annu. Conf. North Amer. Chapter Assoc. Comput. Linguistics (NAACL), 2019, pp. 4171–4186.

[2] Y. Tay, M. Dehghani, D. Bahri, and D. Metzler, "Efficient transformers: A survey," ACM Comput. Surv., vol. 55, no. 6, pp. 1–28, Jul. 2020.

[3] Y. Goldberg, "A primer on neural network mod els for natural language processing," J. Artif. Intell. Res., vol. 57, pp. 345–420, Jul. 2016.

[4] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter," arXiv preprint arXiv:1910.01108, 2019.

[5] X. Jiao, Y. Yin, L. Shang, X. Jiang, X. Chen, L. Li, F. Wang, and Q. Liu, "TinyBERT: Distilling BERT for natural language understanding," arXiv preprint

arXiv:2004.03844, 2020.

[6] J. Frankle and M. Carbin, “The lottery ticket hypothesis: Finding sparse, trainable neural networks,” in Proc. Int. Conf. Learn. Represent. (ICLR), 2019.

[7] R. Zafrir, M. B. Raviv, G. Pereg, and R. Wasserblat, “Q8BERT: Quantized 8-bit BERT,” arXiv preprint arXiv:1910.06188, 2019.

[8] H. Tanaka, A. Kunin, D. L. Yamins, and S. Ganguli, “Pruning neural networks without data,” arXiv preprint arXiv:2006.05467, 2020.

[9] A. Yao, Y. Zhao, D. Wang, Y. Ding, S. Cui, and L. Dai, “ZeroQuant: Efficient post-training quantization for transformers without retraining,” arXiv preprint arXiv:2206.01861, 2022.

[10] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, “ALBERT: A lite BERT for self-supervised learning of language representations,” in Proc. Int. Conf. Learn. Represent. (ICLR), 2020.

[11] Z. Sun, H. Yu, X. Song, R. Liu, Y. Yang, and D. Zhou, “MobileBERT: A compact task-agnostic BERT for resource-limited devices,” in Proc. Annu. Conf. Assoc. Comput. Linguistics (ACL), 2020, pp. 2158–2170.

[12] S. Wang, B. Zhang, Y. Hou, H. Jiang, M. Li, and L. Song, “Linformer: Self-attention with linear complexity,” arXiv preprint arXiv:2006.04768, 2020.

[13] Z. Dai, H. Yang, Y. Yang, J. Carbonell, Q. Le, and R. Salakhutdinov, “Funnel-transformer: Filtering redundant information with progressive downsampling,” in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2020.

[14] R. Kitaev, L. Kaiser, and A. Levskaya, “Reformer: The efficient transformer,” in Proc. Int. Conf. Learn. Represent. (ICLR), 2020.

[15] J. Gou, B. Yu, S. J. Maybank, and D. Tao, “Knowledge distillation: A survey,” Int. J. Comput. Vis., vol. 129, pp. 1789–1819, 2021.

[16] S. Wang, X. Bao, H. Wu, and H. Wang, “MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers,” in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2020.

[17] Z. Sun, H. Yu, X. Song, R. Liu, Y. Yang, and D. Zhou, “Patient knowledge distillation for BERT model compression,” in Proc. Annu. Conf. North Amer. Chapter Assoc. Comput. Linguistics (NAACL), 2019.

[18] A. Liu, Y. Shen, T. Chen, and H. Wu, “EdgeBERT: An efficient BERT adaptation for on-chip inference,” arXiv preprint arXiv:2106.01160, 2021.

[19] S. Wang, Z. Zhang, and B. Liu, “Hardware-aware transformers for efficient NLP,” arXiv preprint arXiv:2007.09269, 2020.

[20] P. Warden and D. Situnayake, TinyML: Machine Learning on Microcontrollers. Sebastopol, CA, USA: O’Reilly Media, 2019.

[21] J. Wu, Y. Zhong, and X. Huang, “FastFormers: Highly efficient transformer models for NLP,” arXiv preprint arXiv:2010.13382, 2020.

[22] E. Strubell, A. Ganesh, and A. McCallum, “Energy and policy considerations for deep learning in NLP,” in Proc. Annu. Conf. Assoc. Comput. Linguistics (ACL), 2019, pp. 3645–3650.

[23] T. Wolf et al., “Transformers: State-of-the-art natural language processing,” in Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP), 2020.

[24] Microsoft Research, “Phi-2: The surprising power of small models,” 2023. [Online]. Available: https://www.microsoft.com/en-us/research/publication/phi-2/

[25] X. Jiang et al., “Mistral-7B: A 7B parameter language model,” 2023. [Online]. Available: https://mistral.ai/news/mistral-7b/

[26] J. Zhang et al., “TinyLlama: An open-source small language model,” 2024. [Online]. Available: https://huggingface.co/TinyLlama

[27] J. Hoffmann et al., “Training compute-optimal large models,” arXiv preprint arXiv:2203.15556, 2022.

[28] R. Bommasani et al., “On the risks of foundation models,” arXiv preprint arXiv:2108.07258, 2021.

[29] W. Fedus et al., “Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity,” in Proc. J. Mach. Learn. Res. (JMLR), 2021.

[30] H. Sutherland et al., “Efficiency metrics in NLP,” arXiv preprint arXiv:2209.11229, 2022