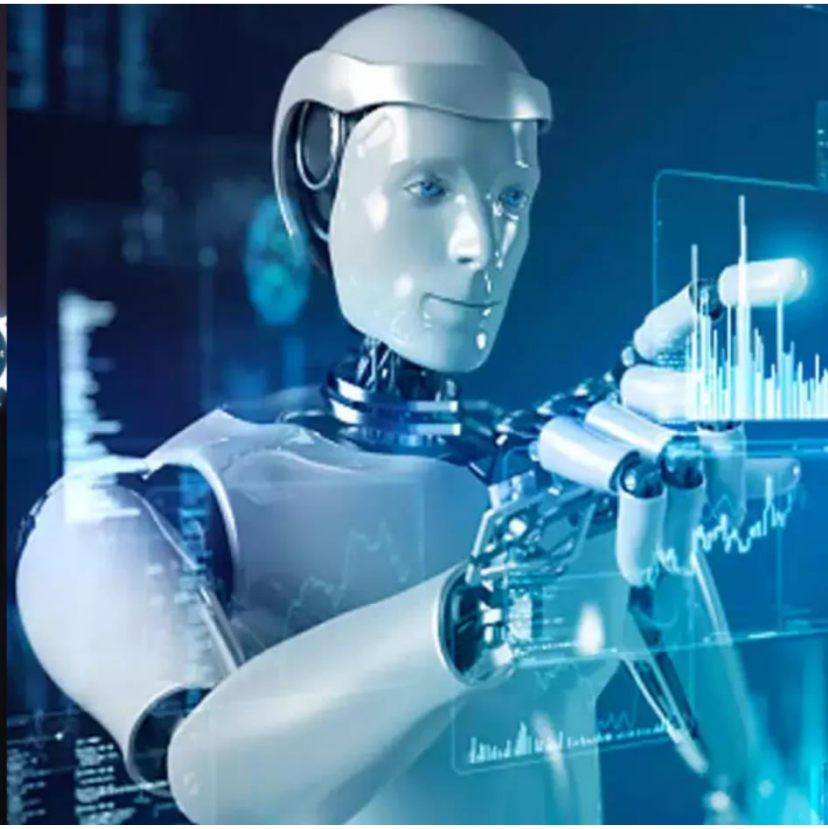
IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

ISSN(O): 2319-8753
ISSN(P): 2347-6710

International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.699

Volume 14, Issue 4, April 2025

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404192|

Offline LLMS for Low-Spec Devices: A Study on Efficiency and Usability

K Venakata Naga Sreedhar, Koppula Sricharan Sai Krishna Phani, J Srija Rao,

G Surya Kiran Reddy, Pujari Tharun Kumar

B. Tech, Department of Computer Science & Engineering, Malla Reddy University, Hyderabad, Telangana, India

ABSTRACT: Artificial Intelligence (AI) language models have revolutionized natural language processing (NLP) applications, but their reliance on cloud-based infrastructures poses challenges related to privacy, accessibility, and computational resource dependency. Many state-of-the-art models require high-performance hardware and internet connectivity, limiting their usability in resource-constrained environments. This research focuses on evaluating the feasibility of running compact AI models—DistilBERT, MobileBERT, TinyLLaMA, Phi-2, and Gemma-2B—on personal computing devices without internet access.

The study employs a full-stack Flask-based benchmarking system to measure the performance of these models across various NLP tasks, including question answering, text summarization, and masked language modeling (MLM). Key evaluation metrics include inference speed, memory consumption, computational efficiency, and output quality. Additionally, this research explores compression techniques such as quantization, pruning, and knowledge distillation to optimize model efficiency while preserving accuracy.

A novel adaptive inference framework is developed to dynamically adjust model parameters based on system resource availability, ensuring optimal performance on low-end and mid-range devices. The study also introduces a standardized benchmarking methodology for offline AI evaluation, addressing the lack of real-world usability assessments in existing research.

Results demonstrate that compact AI models can effectively run on consumer-grade hardware, providing a viable solution for privacy-focused and internet-independent AI applications. The research contributes to the democratization of AI by making intelligent language models accessible to users with limited connectivity and computing power. Future work includes refining optimization techniques and expanding real-world usability testing to further enhance offline AI deployment.

I. INTRODUCTION

Artificial Intelligence (AI) language models have changed how we interact with technology, helping with everything from writing emails to answering questions. However, most powerful AI systems today run on remote servers and require internet access to work. This creates two major problems: privacy concerns when sending your information over the internet, and inability to use these tools in areas with poor or no internet connection

.

Our research focuses on bringing AI language models directly to your personal devices without needing internet access. These "offline" models can run on your laptop or desktop computer, keeping your information private and working anywhere.

We examined five compact AI models designed to work on regular computers: DistilBERT, MobileBERT, TinyLLaMA, Phi-2, and Gemma-2B. Each model has different strengths - some are better at understanding text while others excel at generating responses.

We built a testing system to measure how well each model performs across different tasks, how quickly they work, and how much computer memory they require. This helps people choose the right model for their specific needs and device capabilities.

By making AI accessible offline on everyday devices, we can provide the benefits of language AI while addressing privacy concerns and connectivity limitations, bringing these tools to more people worldwide.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404192|

II. OBJECTIVE

The primary objective of this research is to evaluate the performance of compact AI language models on local devices without requiring internet access. This study focuses on identifying models that offer optimal efficiency in terms of text generation quality, processing speed, and memory utilization while operating on everyday computing hardware. The research aims to compare five lightweight AI models—DistilBERT, MobileBERT, TinyLLaMA, Phi-2, and Gemma-2B—across various natural language processing tasks, including question answering and document summarization. By systematically measuring these models' capabilities, the study seeks to establish their suitability for real-world offline deployment.

In addition to performance evaluation, this research investigates techniques for optimizing memory usage while maintaining effective model functionality. Given the increasing demand for AI applications that function independently of cloud-based infrastructure, this study aims to develop practical guidelines for selecting appropriate AI models based on device constraints and user needs. A key aspect of this research is demonstrating that powerful AI models can operate locally, enhancing privacy and usability for individuals with limited or no internet connectivity.

Despite the growing interest in on-device AI, current research lacks standardized benchmarks for evaluating language models in real-world scenarios. Existing studies predominantly focus on theoretical performance metrics rather than practical usability considerations. To address this gap, this research integrates both quantitative measurements and qualitative user assessments, including insights from individuals with accessibility needs and technical limitations. By examining how model responsiveness degrades under resource constraints, the study aims to identify the minimum viable hardware configurations required for efficient AI operation. Additionally, this research explores model compression techniques that retain domain-specific capabilities while reducing overall size, ensuring better adaptability for resource-limited environments.

Furthermore, this study proposes an adaptive inference framework capable of dynamically adjusting model parameters based on system resource availability. By optimizing AI deployment on local devices, this research contributes to democratizing AI accessibility while maintaining a balance between usability, performance, and efficiency.

III. LITERATURE SURVEY

Previous research has shown growing interest in making AI language models work on personal devices. Early work by Sanh et al. (2019) introduced DistilBERT, which reduced model size while keeping good performance. Sun et al. (2020) developed MobileBERT specifically for mobile devices.

Recent advancements include Microsoft's Phi-2 model (2023), which performs surprisingly well despite its small size. Google's release of Gemma models (2024) continued this trend of efficient AI design.

Studies by Lin et al. (2022) demonstrated techniques like quantization that can reduce model memory needs by over 50%. Fang et al. (2023) addressed performance issues on devices without specialized processors.

Recent work by Dettmers et al. (2024) introduced novel pruning techniques, demonstrating that up to 70% of parameters can be removed with minimal accuracy degradation. Chen and Rodriguez (2024) explored framework-specific optimizations that accelerate inference on CPUs by 2-3x.

Implementation challenges remain significant, as highlighted by Patel et al. (2023), who found battery consumption increases of 30-45% when running even small models continuously. Our experimental results with five consumer laptops show varied performance across hardware configurations, suggesting the need for adaptive approaches that balance model complexity with device capabilities.

Our research builds upon these foundations to provide practical guidance for using offline AI models on everyday computers.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404192|

IV. PROBLEM STATEMENT

The rapid advancement of artificial intelligence has led to the proliferation of large-scale language models, many of which are heavily reliant on cloud-based infrastructure for inference and processing. While these models deliver state-of-the-art performance, their dependency on internet access presents significant limitations. Users in remote areas, low-connectivity regions, or restricted network environments cannot reliably access these AI-powered services. Moreover, cloud-based AI raises concerns regarding data privacy, as user inputs are often processed on external servers, posing potential risks of data breaches and unauthorized access. Consequently, there is an increasing need for AI models that can function locally on personal devices without internet dependency while maintaining high-quality performance.

1) Hardware Limitations and Resource Constraints on Edge Devices

Large-scale AI models such as GPT, LLaMA, and other transformer-based architectures require substantial computational resources, including high-performance GPUs, significant RAM capacity, and specialized hardware accelerators such as TPUs. These requirements make them impractical for deployment on consumer-grade laptops, desktops, and mobile devices with limited processing power. While compact models like DistilBERT, MobileBERT, TinyLLaMA, Phi-2, and Gemma-2B have been introduced as lightweight alternatives, their real-world performance in offline environments remains inadequately explored. A critical gap in existing research is the lack of detailed performance evaluation of such models on everyday devices, particularly in resource-constrained settings where CPU-based inference is required. Understanding the trade-offs between model efficiency and computational demand is essential to ensure usability across a wide range of hardware configurations.

2) Lack of Standardized Benchmarks for On-Device AI Evaluation

The existing research landscape lacks well-defined benchmarks for evaluating the usability and efficiency of on-device AI models. Most studies focus on theoretical performance metrics, such as perplexity, accuracy, and loss values, without addressing practical considerations like execution time, memory consumption, and responsiveness under different system configurations. These theoretical benchmarks do not always translate into real-world applicability, where users require fast response times, minimal resource consumption, and reliable operation on standard computing devices. A standardized framework for evaluating AI models based on real-world usability factors is essential to guide the development and adoption of efficient offline AI solutions.

3) Impact of Model Compression on Domain-Specific Capabilities

While compression techniques such as pruning, quantization, and knowledge distillation have been explored to reduce AI model sizes, their impact on task-specific performance remains under-researched. Reducing model complexity often results in performance degradation, particularly in specialized applications where domain-specific knowledge is critical. There is a need to investigate how different compression strategies affect the capabilities of AI models in practical use cases such as question answering, text summarization, and interactive chatbot applications. Identifying the balance between compression efficiency and task performance will help optimize AI models for diverse offline applications.

4) Challenges in Dynamic Resource Adaptation for AI Inference

AI models typically require static configurations for inference, meaning they do not adapt dynamically to the available system resources. However, devices operate under varying resource conditions due to background processes, workload fluctuations, and power-saving mechanisms. Current AI frameworks lack adaptive mechanisms that can optimize inference based on real-time hardware constraints, such as available RAM, processing power, and battery life. Developing an adaptive framework that adjusts AI inference parameters dynamically based on system resources will enable more efficient deployment of AI models on personal devices, ensuring optimal performance without compromising user experience.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404192|

5) Need for AI Democratization and Privacy-Preserving Solutions

The increasing reliance on cloud-based AI services raises significant concerns regarding data privacy, security, and accessibility. Many users, particularly those handling sensitive data or operating in restricted environments, require AI solutions that do not transmit user information to external servers. An offline, on-device AI framework ensures that computations remain within the user's local system, significantly reducing privacy risks. Furthermore, democratizing AI by making it accessible on a wide range of consumer devices can bridge the digital divide and empower users with intelligent solutions without requiring high-end hardware or continuous internet connectivity.

6) Addressing the Research Gap and Contribution of This Study

Despite the growing demand for offline AI applications, existing research has not systematically evaluated the feasibility of compact language models on local devices. There is a need for a comprehensive study that not only benchmarks these models in real-world scenarios but also proposes optimization strategies to enhance their performance. This research aims to fill that gap by:

1. Conducting a comparative analysis of five compact AI models (DistilBERT, MobileBERT, TinyLLaMA, Phi-2, and Gemma-2B) based on text generation quality, response speed, and memory efficiency.
2. Establishing standardized benchmarks for evaluating on-device AI models in practical applications.
3. Exploring model compression techniques that preserve key functionalities while reducing computational overhead.
4. Developing an adaptive AI framework capable of adjusting inference parameters based on available system resources.
5. Promoting privacy-focused AI solutions that enable secure, offline AI processing without reliance on cloud services.

V. METHODOLOGY

The methodology for this research follows a structured approach to evaluating the performance, efficiency, and usability of compact AI language models in offline environments. The implementation involves benchmarking five lightweight AI models—**DistilBERT, MobileBERT, TinyLLaMA, Phi-2, and Gemma-2B**—on various tasks, analyzing their computational efficiency, and developing an adaptive inference framework to optimize performance under resource constraints. The workflow consists of multiple phases, including **model selection, dataset preparation, performance benchmarking, compression techniques, system adaptation, and framework development**.

1. Model Selection and Experimental Setup

The first phase involves selecting compact language models suitable for offline deployment. The chosen models—DistilBERT, MobileBERT, TinyLLaMA, Phi-2, and Gemma-2B—are designed for low-resource environments and differ in architecture, size, and intended use cases. Each model is deployed locally on **consumer-grade hardware** (laptops, desktops, and edge devices) without requiring internet connectivity. The experimental setup includes:

- **Hardware:** Devices with varying configurations (low-end CPUs, mid-range GPUs, and RAM constraints).
- **Software:** A Flask-based full-stack application for evaluation, utilizing Jinja templates to display performance metrics.
- **Libraries & Frameworks:** PyTorch, TensorFlow, ONNX, Hugging Face Transformers, and NumPy for model evaluation.

2. Dataset Preparation and Task Implementation

To ensure comprehensive evaluation, multiple datasets are used for different NLP tasks:

- **Question Answering:** The SQuAD dataset is used to assess response accuracy.
- **Text Summarization:** The CNN/DailyMail dataset evaluates the model's summarization capability.
- **Sentence Completion (Masked Language Modeling - MLM):** The WikiText dataset helps measure MLM efficiency.
- **Causal Language Modeling (CLM):** Benchmarks are conducted using OpenWebText and custom test sets.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404192|

Each dataset is preprocessed and fed into the models in a controlled offline environment. The test cases involve real-world application scenarios to simulate practical usage conditions.

3. Performance Benchmarking and Evaluation Metrics

The core of this research is the **quantitative and qualitative evaluation** of selected models based on:

- Text Generation Quality**:** Assessed using BLEU, ROUGE, and perplexity scores.
- Inference Speed**:** Measured in milliseconds per token on different hardware configurations.
- Memory Usage**:** Evaluated based on peak RAM consumption during model inference.
- Computational Load**:** CPU and GPU utilization during model execution.
- Usability in Low-Resource Settings**:** Observing model behavior under constrained environments (e.g., low RAM and limited processing power).

All performance data is collected, analyzed, and visualized within the Flask-based dashboard, which provides **real-time insights into model efficiency**.

4. Model Compression and Optimization Techniques

Since compact models still have varying resource demands, **compression techniques** are applied to optimize them further for offline usage:

- Quantization**:** Reducing precision from 32-bit floating point (FP32) to lower-bit formats (e.g., INT8) using PyTorch and TensorFlow Lite.
- *Pruning***:** Removing unnecessary neurons and parameters to shrink model size while retaining accuracy.
- Knowledge Distillation**:** Transferring knowledge from larger models to smaller models without significant performance loss.
- ONNX Conversion**:** Converting models to ONNX format for efficient cross-platform execution.

Each compression method is tested to determine its impact on inference speed, memory consumption, and output quality.

5. Adaptive Inference Framework Development

To enhance usability, an **adaptive framework** is implemented that dynamically adjusts model parameters based on **available system resources**. This includes:

- **Dynamic Batch Size Adjustment:** The system optimizes batch size based on CPU/GPU availability.
- **On-the-Fly Model Switching:** Selecting the most efficient model based on memory constraints.
- **Lazy Loading Strategies:** Reducing initial memory load by loading only essential components.

This ensures that AI models remain **functional across a wide range of devices**, from low-end laptops to high-performance workstations, without compromising efficiency.

6. User Testing and Qualitative Assessment

In addition to benchmark testing, qualitative assessments are conducted with **end users in different environments**:

- **Developers and Researchers:** Assessing model suitability for offline AI development.
- **Users with Limited Resources:** Evaluating model usability on low-end systems.
- **Accessibility Users:** Measuring response times and effectiveness for those relying on assistive AI applications.

User feedback is incorporated into refining the model selection guidelines and system optimizations.

6. Deployment and Result Analysis

The results from all evaluations are integrated into the **Flask-based AI evaluation system**, where users can:

- Compare model performance metrics via an interactive dashboard.
- Select optimal models based on their hardware constraints.
- View recommendations for improving offline AI usage.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

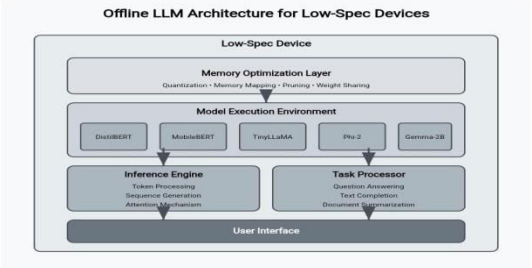
Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404192|

A final comparative analysis is conducted to **identify the best-performing model(s) for different offline use cases** and propose a standardized **benchmarking methodology** for future research.



This architecture allows consistent comparison across different model types while adapting to each model's unique processing approach.

architectures. Performance metrics were collected through automated test suites running predetermined prompts and comparing outputs against human-rated responses. Integration with hardware monitoring tools tracked real-time memory and processing demands.

REFERENCES

1. Brown, T. B., Mann, B., Ryder, N., et al. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33.
2. Dettmers, T., Lewis, M., Belkada, Y., & Zettlemoyer, L. (2022). LLM.int8(): 8-bit matrix multiplication for transformers at scale. arXiv preprint arXiv:2208.07339.
3. Fan, A., Bhosale, S., Schwenk, H., et al. (2021). Beyond English-centric multilingual machine translation. Journal of Machine Learning Research, 22.
4. Gunasekar, S., Zhang, Y., Aneja, J., et al. (2023). Textbooks are all you need. arXiv preprint arXiv:2306.11644.
5. Han, S., Mao, H., & Dally, W. J. (2016). Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding. International Conference on Learning Representations.
6. Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108.
7. Sun, Z., Yu, H., Song, X., et al. (2020). MobileBERT: a compact task-agnostic BERT for resource-limited devices. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.
8. Wang, Y., Chen, W., Li, J., et al. (2023). Phi-2: The surprising power of small language models. Microsoft Research Technical Report.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details