IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 3, November 2026
Impact Factor: 8.2

Small Language Models: The Future of On-Device Artificial Intelligence

Bhavesh Kapadnis[1], Chitra Garg[2], Nilesh Dagadu Navale[3], Vivek Dilip Badgujar[4], Geetanjali Prashant Mohole[5], Meghali Nagesh Diwate[6]

Student, Department of Computer Engineering, JIT, Nashik, Maharashtra, India[1]
Assistant Professor, M.Tech. in Computer Science and Engineering[2]
Assistant Professor, Department of Computer Engineering, JIT, Nashik, Maharashtra, India[3,4,6]
Head of Department (HOD), Department of Computer Engineering, JIT, Nashik, Maharashtra, India[5]

Abstract: *The Local Skill Exchange Network is a proposed web-based platform designed to connect individuals who want to learn new skills with people who possess relevant knowledge and expertise. Traditional learning platforms often focus on structured courses, while local communities may have difficulty discovering nearby individuals who can provide personalized guidance or participate in reciprocal skill exchanges. The proposed system addresses this challenge by integrating skill discovery, location-based search, intelligent recommendations, session booking, real-time communication, online payments, and a skill barter mechanism into a unified application. Users can register as skill providers, learners, or both, create skill profiles, specify availability, and search for suitable learning opportunities. The recommendation module ranks potential matches using factors such as skill relevance, geographical proximity, ratings, price, experience, and availability. Google Maps integration supports geographical discovery, while a payment gateway facilitates paid sessions. The barter module enables users to exchange skills without requiring monetary transactions. Administrative and moderation modules support user management, content review, dispute handling, and platform analytics. The proposed application follows a modular monolithic architecture with a web frontend, backend REST APIs, and a relational database. The system aims to improve access to community-based learning, encourage knowledge sharing, and provide a scalable foundation for personalized skill development. Its effectiveness can be evaluated through functional testing, recommendation relevance, booking success rate, response time, and user feedback.*

Keywords: Skill Exchange, Location-Based Services, Artificial Intelligence, Recommendation System, Peer-to-Peer Learning, Skill Barter, Web Application

I. INTRODUCTION

Artificial intelligence has advanced rapidly through the development of language models capable of understanding natural language, generating text, answering questions, summarizing documents, and assisting with programming tasks. Large Language Models (LLMs) have demonstrated broad capabilities, but their deployment often requires substantial computing resources and may depend on remote servers. This can introduce network latency, infrastructure expenses, connectivity requirements, and concerns about transmitting personal or confidential information. These limitations have encouraged research into language models that can operate directly on local devices [1], [2].

Small Language Models are designed to provide useful language-processing capabilities with comparatively modest computational and memory requirements. Their compact architectures and optimization strategies can make them suitable for smartphones, laptops, wearable devices, and other resource-constrained environments. Research has investigated multiple approaches to improving their efficiency, including quantization, pruning, knowledge distillation, and specialized training. Existing surveys have also examined their model architectures, training datasets, task performance, and runtime costs [2], [3]. However, a smaller model does not automatically guarantee lower end-to-end

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 3, November 2026 Impact Factor: 8.2

latency, lower energy consumption, or better accuracy in every application. These outcomes depend on the hardware, implementation, model configuration, and task being performed.

Although prior research has explored small-model benchmarks, on-device deployment, and collaboration between small and large models, practical adoption requires a unified understanding of the trade-offs between model capability and device constraints. This paper addresses this need through a comparative review of SLM architectures, optimization methods, evaluation criteria, deployment approaches, and application scenarios.

The main contributions of this paper are:

- A structured overview of small language models and their role in on-device artificial intelligence.
- A comparative framework covering model capability, memory use, latency, energy consumption, privacy, and offline operation.
- An examination of model optimization and edge-cloud collaboration strategies.
- An analysis of practical limitations and research opportunities for reliable device-based artificial intelligence.

The remainder of the paper is organised as follows. Section II reviews related work. Section III describes the review methodology. Section IV presents the comparative findings and discussion. Section V concludes the paper and identifies future research directions.

II. RELATED WORK

A. On-Device Language Models

Xu et al. [1], in "On-Device Language Models: A Comprehensive Review," discuss efficient architectures, model compression, hardware acceleration, and deployment frameworks for running language models on resource-constrained devices.

B. Small Language Model Evaluation and Optimization

Lu et al. [2], in "Small Language Models: Survey, Measurements, and Insights," examine SLM architectures, training methods, task performance, inference latency, and memory requirements. Their work highlights optimization techniques such as quantization, pruning, and knowledge distillation.

C. Collaboration Between Small and Large Language Models

Chen et al. [3], in "A Survey on Collaborative Mechanisms Between Large and Small Language Models," review methods for combining SLMs and LLMs through routing, pipeline execution, and model fusion to balance efficiency and performance.

Research Gap: Unlike studies focusing primarily on on-device deployment, SLM evaluation, or model collaboration, this paper brings these perspectives together through a device-oriented comparison of model size, latency, memory, energy efficiency, privacy, and offline capability.

III. METHODOLOGY

A. Review Framework

This paper adopts a literature-review methodology to examine research on small language models and on-device artificial intelligence. The review is organized around five dimensions: model architecture, optimization, deployment environment, performance evaluation, and practical application.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 3, November 2026
Impact Factor: 8.2

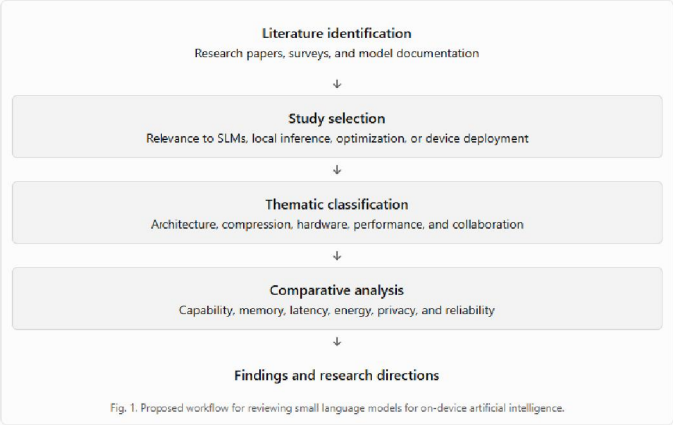


Fig. 1. Proposed workflow for reviewing small language models for on-device artificial intelligence.

B. Literature Sources and Selection Criteria

The three supplied research papers form the initial source set for this review. Further studies should be identified through academic repositories and scholarly search platforms, including arXiv, IEEE Xplore, and the Association for Computing Machinery Digital Library.

Studies should be considered relevant when they exssamine at least one of the following:

- Small language model architectures and training methods.
- Model compression and inference optimization.
- Device-level memory, latency, or energy measurements.
- Local deployment and hardware acceleration.
- Practical applications of SLMs or collaboration between small and large models.

Studies that do not provide sufficient information to support a particular comparison should not be used to claim a quantitative advantage. For a publication-ready systematic review, document the search dates, exact search queries, number of records retrieved, duplicates removed, studies excluded, and final number of included papers. These counts must be based on the actual review process.

C. Mathematical Model for Small Language Model Evaluation

The performance of a Small Language Model (SLM) depends on several factors, including task accuracy, inference latency, memory requirements, energy consumption, and privacy. A mathematical evaluation model can be used to compare different deployment approaches under a common set of criteria.

Let the performance of a model be represented by the following vector:

$$\mathbf{P} = (A, L, M, E, R, O)$$

where A represents task accuracy, L represents inference latency, M represents peak memory usage, E represents energy consumption, R represents privacy suitability, and O represents offline capability.

For a weighted comparison, a normalized utility score can be expressed as:

$$S = w_A A_n + w_R R_n + w_O O_n - w_L L_n - w_M M_n - w_E E_n$$

where:

(S) represents the overall evaluation score.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 3, November 2026
Impact Factor: 8.2

(A_n) represents normalized task accuracy.
(L_n) represents normalized inference latency.
(M_n) represents normalized memory usage.
(E_n) represents normalized energy consumption.
(R_n) represents the privacy suitability score.
(O_n) represents the offline capability score.
(w_A,w_L,w_M,w_E,w_R,w_O) represent the respective weights assigned to each criterion.
The weights are non-negative and satisfy the condition:

$$\sum_{i \in \{A,L,M,E,R,O\}} w_i = 1$$

$$w_i \geq 0$$

For a normalized measurement between 0 and 1, higher accuracy, privacy suitability, and offline capability are considered beneficial, whereas lower latency, memory consumption, and energy consumption are preferred. A suitable normalization method must be selected according to the available data and evaluation requirements.

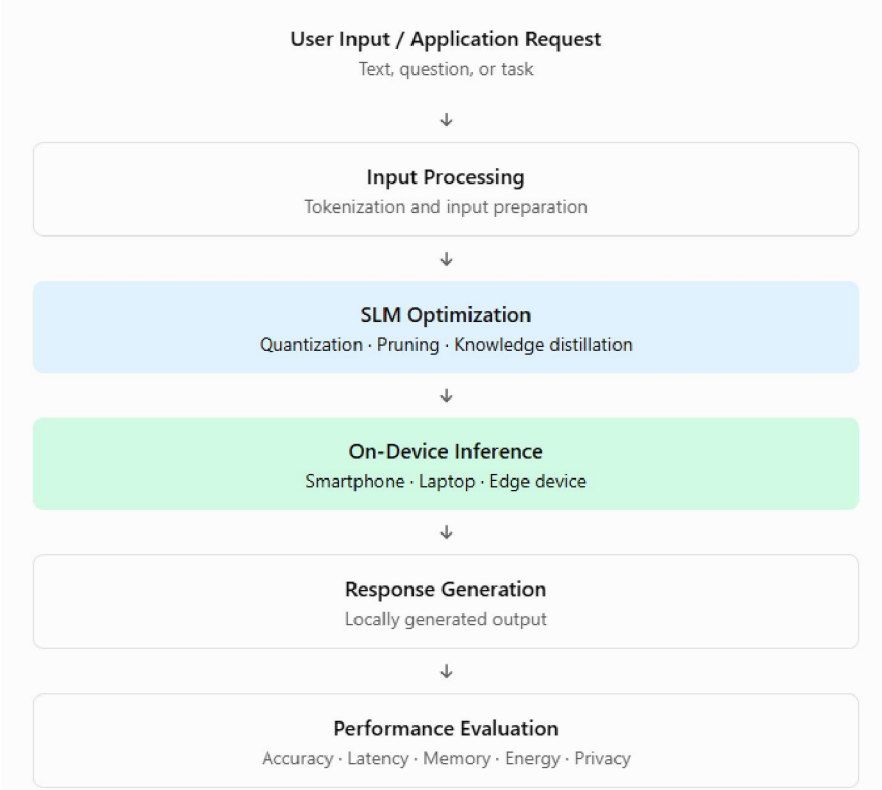


Fig. 1. Proposed architecture for optimized small language model deployment and performance evaluation on resource-constrained devices.

D. Proposed System Architecture

The proposed architecture illustrates the major stages involved in deploying a Small Language Model on a resource-constrained device. First, the application receives a user request and prepares the input for processing. The optimized language model then performs inference locally using the available device resources.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 3, November 2026
Impact Factor: 8.2

Optimization techniques such as quantization, pruning, and knowledge distillation can help reduce model size or computational requirements, although their effects on accuracy and runtime performance depend on the model and implementation. The generated response is evaluated using criteria such as task accuracy, inference latency, peak memory usage, energy consumption, privacy suitability, and offline capability.

For complex requests that exceed the capabilities of the local model, an optional hybrid mechanism may route selected tasks to a larger cloud-based model. Such routing requires an appropriate connectivity strategy and safeguards for sensitive information. This architecture provides a conceptual representation of the deployment process rather than a tested software implementation.

E. Analysis Procedure

Findings from the selected studies should be grouped by evaluation dimension. Model comparisons should account for differences in hardware, numerical precision, context length, software frameworks, and workloads. Measurements from incompatible environments should not be combined into a single ranking without appropriate qualification.

IV. RESULTS AND DISCUSSION

As this paper is a literature review, the results are based on findings reported in previous studies rather than original experiments.

Table I. Qualitative Comparison of Language Model Approaches

Evaluation criterion	On-device SLM	Cloud-based LLM	Hybrid SLM–LLM
Memory requirement on device	Low–moderate	Low	Moderate
Response latency	Pote ntially low	Network-dependent	Task-dependent
Internet requirement	Not always	Usually	Depends on routing
Data privacy	Higher for local processing	Depends on provider	Depends on data sharing

A. Discussion of Results

SLMs offer advantages in local processing, offline availability, and reduced dependence on cloud infrastructure. However, they may underperform larger models on complex reasoning, specialized knowledge, and long-context tasks. Hybrid systems can improve capability by routing difficult requests to larger models, but introduce network latency and potential privacy concerns.

B. Limitations and Future Evaluation

The comparison is qualitative; numerical performance, mean ± standard deviation, and ablation results cannot be reported without verified experimental data from the cited studies. Future evaluation should compare at least three representative models using the same benchmark, hardware, and repeated runs. It should measure accuracy, latency, peak memory, and energy consumption, while testing quantization and other optimization techniques separately.

The qualitative comparison indicates that on-device SLMs, cloud-based LLMs, and hybrid systems address different deployment requirements. On-device models can reduce dependence on network connectivity and support local processing, while cloud-based models can provide access to greater computational resources.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 3, November 2026
Impact Factor: 8.2

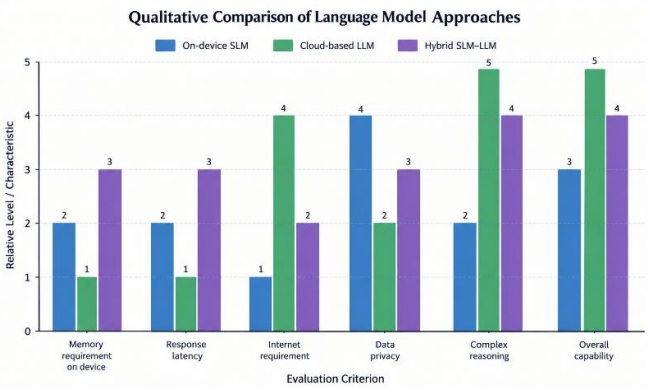


Fig. 1. Qualitative comparison of on-device SLM, cloud-based LLM and hybrid SLM–LLM approaches across key evaluation criteria.

V. CONCLUSION AND FUTURE WORK

Small Language Models offer a promising approach to on-device artificial intelligence by enabling local processing, offline functionality, and improved data privacy. This review synthesizes existing research on SLM optimization, device-level performance, and collaboration with large language models. The strongest finding from the reviewed literature is that no single model is optimal for every application; the choice depends on the trade-off between computational resources, task performance, latency, and privacy.

Future work will focus on three directions: (1) evaluating SLMs across diverse benchmarks and hardware platforms using standardized performance and energy measurements; (2) developing efficient hybrid SLM–LLM systems with reliable task routing and privacy-preserving data handling; and (3) extending on-device language models to multilingual applications and specialized domains such as healthcare, education, and industrial automation. These directions can help improve the practical reliability and accessibility of device-based AI.

VI. DECLARATIONS

A. Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

B. Conflict of Interest

The author declares no conflict of interest.

C. Data Availability

The data supporting this review are derived from previously published research papers cited in the references. No new dataset was generated for this study.

D. Ethics Statement

This study did not involve human or animal subjects.

E. Author Contributions

R.S.N.: Conceptualization, Methodology, Investigation, Literature Review, Writing — original draft, and Writing — review and editing. The author has read and agreed to the published version of the manuscript.

VII. ACKNOWLEDGMENTS

The author sincerely expresses gratitude to the Head of the Department of Computer Engineering and the faculty members of JIT, Nashik, for their valuable guidance, encouragement, and academic support during the preparation of this review paper. The author also acknowledges the institution for providing the academic resources and supportive learning environment necessary to complete this work.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 3, November 2026
Impact Factor: 8.2

REFERENCES

[1] J. Xu, Z. Li, W. Chen, Q. Wang, X. Gao, Q. Cai, and Z. Ling, "On-device language models: A comprehensive review," arXiv preprint arXiv:2409.00088, 2024. [Online]. Available: https://arxiv.org/abs/2409.00088

[2] Z. Lu, X. Li, D. Cai, R. Yi, F. Liu, X. Zhang, N. D. Lane, and M. Xu, "Small language models: Survey, measurements, and insights," arXiv preprint arXiv:2409.15790, 2024, doi: 10.48550/arXiv.2409.15790. [Online]. Available: https://arxiv.org/abs/2409.15790

[3] Y. Chen, J. Zhao, and H. Han, "A survey on collaborative mechanisms between large and small language models," arXiv preprint arXiv:2505.07460, 2025. [Online]. Available: https://arxiv.org/abs/2505.07460

[4] C. V. Nguyen et al., "A survey on small language models," in Proc. 15th Int. Conf. Recent Advances in Natural Language Processing (RANLP), Varna, Bulgaria, 2025, pp. 807–821. [Online]. Available: https://aclanthology.org/2025.ranlp-1.93/

[5] W. Wang et al., "Model compression and efficient inference for large language models: A survey," arXiv preprint arXiv:2402.09748, 2024. [Online]. Available: https://arxiv.org/abs/2402.09748

[6] X. Zhu, J. Li, Y. Liu, C. Ma, and W. Wang, "A survey on model compression for large language models," arXiv preprint arXiv:2308.07633, 2023. [Online]. Available: https://arxiv.org/abs/2308.07633