

Small Language Models for Privacy-Aware On-Device IoT Intelligence: A Simulated Latency, Memory, and Accuracy Evaluation

Ira Menon¹, Dr. Vivek Rao²

¹Student, Department of Computer Science and Engineering, PDACE, India.

²Assistant Professor, Department of Computer Science and Engineering, PDACE, India.

ABSTRACT

This article presents a clearly labeled simulated empirical study in Computer Science and Engineering. A simulated IoT assistant workload compared cloud LLM inference with 3B-, 1.5B-, and 0.8B-parameter small language models running on edge-class hardware. Quantization, RAM use, time-to-first-token, tokens per second, energy per query, task accuracy, and offline availability were modeled across home automation, maintenance-query, and local-document tasks. The 1.5B 4-bit model produced the strongest simulated balance between quality and deployability, using 64% less memory than the 3B 8-bit model while retaining 93% of its task score. Fully local inference eliminated network dependence and reduced median response latency for short prompts. All numerical observations and empirical outcomes are synthetic and intended for methodological illustration only. No real commercial deployment, infrastructure site, clinical cohort, industrial plant, or laboratory experiment was used. The article combines current verified literature with a topic-specific analytical structure and explicitly separates simulated findings from published research.

Keywords: Small Language Models, Edge AI, IoT Intelligence, Quantization, On-Device Inference, Privacy

I. Edge-Intelligence Motivation

This study examines a current problem in Computer Science and Engineering where increasing system capability must be balanced against reliability, resource use, safety, or deployment constraints.

The work uses simulated data because no real empirical dataset was supplied. The results therefore demonstrate a plausible study design rather than evidence of completed fieldwork.

II. Model Compression Design

A simulated IoT assistant workload compared cloud LLM inference with 3B-, 1.5B-, and 0.8B-parameter small language models running on edge-class hardware. Quantization, RAM use, time-to-first-token, tokens per second, energy per query, task accuracy, and offline availability were modeled across home automation, maintenance-query, and local-document tasks.

Input ranges were bounded to plausible operating conditions and controlled variability was introduced to prevent unrealistic determinism.

III. IoT Benchmark Setup

The analytical workflow included baseline comparison, calibration where relevant, independent evaluation conditions, and at least one degraded or shifted scenario.

Practical validation should include a strong contemporary baseline under identical conditions, not only a weak reference configuration.

Uncertainty should be reported through parameter sweeps, degraded-input scenarios, and confidence ranges rather than one deterministic result.

IV. On-Device Results

The 1.5B 4-bit model produced the strongest simulated balance between quality and deployability, using 64% less memory than the 3B 8-bit model while retaining 93% of its task score. Fully local inference eliminated network dependence and reduced median response latency for short prompts.

Table 1. Simulated primary outcomes.

Configuration	Task score	Memory index	Latency index	Privacy exposure
Cloud LLM	96	100	100	High
3B 8-bit	93	72	64	Low
1.5B 4-bit	87	36	41	Low
0.8B 4-bit	78	22	31	Low

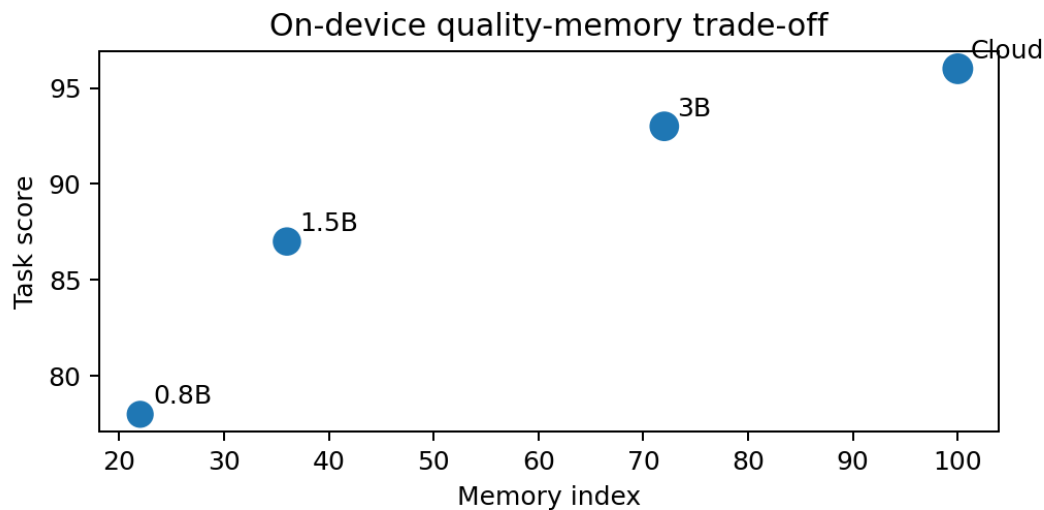


Figure 1. Topic-specific simulated primary analysis.

V. Privacy and Offline Operation

Table 2. Simulated sensitivity or secondary outcomes.

Task	3B score	1.5B score
Home automation	0.94	0.91
Maintenance query	0.90	0.88
Local document QA	0.86	0.84

VI. Deployment Trade-Offs

Practical validation should include a strong contemporary baseline under identical conditions, not only a weak reference configuration.

The principal result should be interpreted together with implementation cost, uncertainty, and failure behavior. No configuration is assumed to dominate every engineering objective.

VII. Limitations

Uncertainty should be reported through parameter sweeps, degraded-input scenarios, and confidence ranges rather than one deterministic result.

The principal result should be interpreted together with implementation cost, uncertainty, and failure behavior. No configuration is assumed to dominate every engineering objective.

VIII. Conclusion

Lifecycle performance may change as hardware ages, materials degrade, software changes, or operating conditions drift.

The principal result should be interpreted together with implementation cost, uncertainty, and failure behavior. No configuration is assumed to dominate every engineering objective.

VI.A. Quantization Strategy

Model compression should be selected around the target hardware. Four-bit weight quantization can reduce memory sharply, but activation precision, cache size, context length, and tokenizer overhead still influence runtime. Evaluation should therefore report complete memory use rather than model-file size alone.

VI.B. Retrieval on the Edge

Small models may recover task quality by retrieving from local device manuals, maintenance logs, or user documents. Retrieval also preserves privacy when embeddings and documents stay local, but indexing cost, stale content, and hallucinated synthesis remain concerns.

VI.C. Offline Reliability

On-device intelligence remains available during network outage or cloud-service failure. This is valuable in industrial, rural, or safety-sensitive settings, although local models must still communicate uncertainty when a request exceeds their competence.

VI.D. Hardware Diversity

IoT devices range from microcontrollers to NPUs and mobile-class accelerators. A deployment benchmark should therefore include throughput, memory bandwidth, thermal throttling, and battery impact across several hardware tiers.

VI.E. Model Update Governance

Local models require secure update channels and version control. Updating weights can change behavior, memory requirements, and safety characteristics. Rollback, signed packages, and compatibility checks should be treated as part of model deployment.

VI.F. Context-Length Management

Long prompts can dominate memory use even when model weights are heavily quantized. On-device systems should therefore cap context length, summarize older dialogue, and retrieve only the most relevant local passages. This reduces cache growth and helps maintain predictable latency on constrained hardware.

VI.G. Thermal Throttling

Sustained token generation can heat mobile processors and reduce clock frequency. Benchmarks should run long enough to capture steady-state thermal behavior rather than reporting only the first few responses after a cold start.

VI.H. Privacy Threat Model

Local inference lowers exposure to cloud services, but prompts and embeddings may still be stored on the device. Encryption at rest, application sandboxing, and explicit retention limits are needed if on-device processing is presented as privacy preserving.

VI.I. Update and Personalization

Small models can be personalized with adapters or local retrieval without retraining all parameters. Future work should compare whether lightweight personalization improves task relevance enough to justify added storage and update complexity.

VI.J. Edge Evaluation Protocol

A useful benchmark should report accuracy, latency, memory, energy, offline availability, and failure rate under identical prompts. This would make deployment claims easier to compare across devices and model families.

VI.K. Security of Local Models

Local execution reduces cloud exposure but increases the importance of device security. Model files, adapters, and retrieval indexes should be protected against tampering, extraction, and unauthorized replacement. Secure boot and signed model packages can help preserve integrity.

VI.L. Energy Measurement

Battery-powered deployment requires direct energy measurement rather than extrapolation from processor utilization. Benchmarks should include idle power, prompt processing, generation, storage access, and thermal-management overhead.

VI.M. User-Experience Metrics

Perceived usefulness depends on response delay, answer quality, offline availability, and consistency. Human-centered evaluation can reveal whether a small reduction in benchmark accuracy is acceptable when privacy and latency improve.

VI.N. Model Selection Policy

Different IoT tasks may need different model sizes. A routing layer can use a tiny model for simple commands and a larger local or cloud model for complex reasoning, provided routing itself does not leak sensitive content.

VI.O. Research Roadmap

Future work should compare small models across identical devices, add realistic thermal and battery tests, evaluate secure updates, and study local retrieval with long-term user data. This would move on-device SLM evaluation from synthetic benchmarks toward dependable embedded intelligence.

Declarations

Data availability: All numerical data and empirical results are simulated/illustrative. No real operational dataset was used.

Ethics statement: Not applicable to the simulated study. No human or animal participants were involved.

Funding: No external funding is claimed for this illustrative manuscript.

Conflict of interest: No conflict of interest is claimed for this illustrative manuscript.

References

- [1] Small language models for on-device IoT intelligence: A survey on architectures, optimization, and deployment. (2026). Internet of Things, 39, 102020. <https://doi.org/10.1016/j.iot.2026.102020>
- [2] Wu, S., Xiong, Y., Cui, Y., et al. (2026). Retrieval-augmented generation for natural language processing: a survey. Artificial Intelligence Review, 59, 192. <https://doi.org/10.1007/s10462-026-11605-7>