
DOCTORAL RESEARCH PROPOSAL

**Optimizing Small-Scale Generative AI
Models for Low-Resource Edge Devices
Using Lightweight Compression
Techniques**

Prepared by:

Malak Aliouche

aliouchemalak706@gmail.com

+213 673 633 136

June, 2026

Abstract

This doctoral research proposes a practical framework for deploying small-scale generative AI models on low-resource edge devices. The core problem addressed is the computational incompatibility between generative models and devices with limited memory and no dedicated graphics processor.

Unlike existing research that focuses on large language models requiring expensive hardware [1, 2, 3], this work explicitly targets resource-constrained researchers. All experiments are designed to run on standard consumer laptops using free or low-cost cloud services. The proposed methodology combines three lightweight compression techniques: unstructured pruning, post-training quantization, and self-distillation, drawing from recent advances in tiny machine learning for edge intelligence [4, 5].

Expected outcomes include significant compression ratios with acceptable accuracy retention on standard benchmarks. All code, compressed models, and experimental results will be open-sourced, enabling reproduction by researchers worldwide with minimal budgets.

Keywords: Edge AI, Model Compression, Small-Scale Generative Models, Low-Resource Computing, Reproducible Research

1 Introduction

Generative artificial intelligence has witnessed remarkable advances in recent years. Models such as GPT, BERT, and their variants have achieved state-of-the-art performance in natural language understanding and generation tasks. These models have transformed how researchers approach language-related problems.

However, these impressive capabilities come at a significant computational cost. Even relatively small models require substantial memory, while typical edge devices such as smartphones and Raspberry Pi have limited resources and no dedicated graphics processing unit. This creates a fundamental barrier to deploying generative AI locally [6].

The problem becomes more acute in developing countries, where access to high-end computing hardware is severely limited. Many universities operate with standard consumer laptops with moderate memory and integrated graphics only. Cloud-based AI raises concerns regarding data privacy, network latency, and energy consumption [8].

The present research proposes a lightweight compression framework designed specifically for researchers with limited hardware resources. Unlike most existing work that targets large language models requiring expensive GPU clusters [1, 2, 3], this research focuses exclusively on small-scale generative models and uses only compression techniques that can run on standard consumer hardware.

2 Problem Statement

Most existing compression research suffers from two fundamental problems that make it inaccessible to researchers with limited hardware resources.

The first problem is excessive hardware requirements. Studies using large language models require expensive GPU clusters costing thousands of dollars [3], placing them beyond the budget of individual researchers and small laboratories.

The second problem is lack of reproducibility. Even when research targets smaller models, many papers provide code that cannot run on standard laptops or free cloud services, requiring significant technical expertise and additional investment.

This research directly addresses this gap by designing a compression framework that works entirely within the constraints of standard consumer hardware, using free cloud services only as optional supplements. Recent work in tiny machine learning for embedded systems [5, 4] has demonstrated the feasibility of such an approach.

3 Research Questions

This research seeks to answer the following questions.

Research Question 1: What is the maximum achievable compression ratio for small-scale generative models using lightweight techniques that run on standard consumer hardware?

Research Question 2: How does the order of applying pruning, quantization, and distillation affect generation quality when computational resources are limited?

Research Question 3: Can standard consumer hardware combined with free cloud services reproduce compression results reported in high-end GPU research?

4 Literature Review

4.1 Pruning Techniques

Han and colleagues introduced deep compression, demonstrating that neural networks can be significantly pruned without substantial accuracy loss. Subsequent work applied pruning to BERT models, achieving smaller variants. Comprehensive surveys on model compression techniques [1, 7] highlight pruning as one of the most effective structural optimization methods for language models. However, such work has primarily focused on discriminative rather than generative models.

4.2 Quantization Methods

Gholami and colleagues provided a comprehensive survey of quantization methods. Post-training quantization requires minimal computational overhead and can be performed on standard hardware, making it attractive for low-resource researchers [2]. However, quantization can affect generation quality, especially for longer texts. Recent advances in edge-cloud collaborative compression [8] have shown promising results for bandwidth-efficient deployment.

4.3 Knowledge Distillation

Hinton and colleagues proposed knowledge distillation, where a smaller student model learns from a larger teacher. Self-distillation, where the same model serves as both teacher and student, offers a lightweight alternative requiring no additional hardware. The survey by Gladys et al. [7] provides an extensive review of distillation techniques for building expert small models.

4.4 Tiny Machine Learning for Edge Devices

Recent research has increasingly focused on deploying machine learning on resource-constrained hardware [6]. Alozie et al. [4] and El Mrabet et al. [5] have demonstrated that tiny machine learning (TinyML) can enable AI applications on microcontrollers and IoT devices, providing a foundation for the present work.

4.5 Research Gaps

Four gaps emerge from this review. First, most compression research targets large models requiring expensive hardware [1, 2, 3]. Second, existing work rarely considers the specific challenges of generative models. Third, there is a lack of reproducible benchmarks for consumer hardware. Fourth, few studies have systematically evaluated the combined effect of pruning, quantization, and distillation order on generation quality.

5 Proposed Methodology

The research will be conducted in four sequential phases.

5.1 Phase One: Baseline Establishment

Baseline metrics will be established for selected small-scale generative models on target hardware platforms, including edge devices and standard laptops. Metrics will include inference latency, memory usage, perplexity, and generation quality scores. The methodology follows best practices identified in systematic reviews of model compression [3].

5.2 Phase Two: Compression Implementation

Three compression techniques will be implemented sequentially, drawing from established literature [7, 2]. First, magnitude-based pruning at various levels, followed by fine-tuning. Second, post-training quantization using native libraries. Third, self-distillation for additional quality recovery. Data importance driven compression techniques [8] will also be explored.

5.3 Phase Three: Evaluation

Evaluation will be conducted on standard benchmarks including GLUE, SQuAD, and CNN/DailyMail, plus custom tests for generation quality. Multiple combinations of techniques will be compared, with results reported using appropriate statistical measures. Performance will be assessed on both edge devices and standard laptops following TinyML evaluation protocols [4, 5].

5.4 Phase Four: Dissemination

Research results will be submitted to peer-reviewed journals. All code will be open-sourced with complete documentation to ensure reproducibility.

6 Benchmarking Environment and Evaluation Criteria

6.1 Hardware Platforms

Target hardware includes low-resource edge devices such as Raspberry Pi (similar to setups reviewed in biological sensing applications [6]), standard consumer laptops, and free cloud computing tiers. This range ensures the research remains accessible to researchers with limited budgets.

6.2 Datasets and Metrics

Standard datasets will be used, including GLUE for text classification, SQuAD for question answering, and CNN/DailyMail for summarization. Custom tests will assess generation coherence. Evaluation metrics include compression ratio, memory usage, inference speed, and generation quality measures. These metrics align with those recommended in comprehensive surveys [1, 2].

7 Hardware Constraints and Scope

This research is designed for standard consumer hardware without dedicated graphics processors. Large language models exceeding practical memory limits are explicitly excluded, as the

goal is reproducible research accessible to low-resource researchers. This approach is consistent with edge AI principles outlined by Alozie et al. [4].

8 Risk Analysis

Potential risks include cloud service interruptions, compression-related quality degradation, and hardware unavailability. Mitigation strategies include frequent checkpointing, early stopping mechanisms, and emulation fallbacks. Transparency about limitations will be maintained throughout.

9 Ethical Statement

This research involves no human participants or personal data. All datasets are publicly available. Environmental impact will be minimized through carbon-aware scheduling and use of pre-trained models. All outputs will be openly licensed.

10 Expected Outcomes

This research will produce an open-source compression library suitable for low-resource hardware, a complete reproduction package with documentation, empirical benchmark results across multiple models and platforms, peer-reviewed publications, and educational materials for resource-limited settings.

11 Timeline

The research will be carried out over the duration of the doctoral program. The initial period focuses on literature review and baseline establishment. The middle period is dedicated to implementing compression techniques and conducting evaluations. The final period focuses on analysis, writing, and dissemination of results.

Activity	Year 1	Year 2	Year 3
Literature review & baseline	X		
Compression implementation	X	X	
Evaluation & analysis		X	X
Paper writing & thesis			X

Table 1: Summary of research activities by year

Milestones

Key milestones include completion of baseline results, implementation of the full compression pipeline, completion of evaluation, submission of research papers, and final thesis submission.

References

- [1] Dantas, P. V., Cordeiro, L. C., & Junior, W. S. S. (2025). A review of state-of-the-art techniques for large language model compression. *Complex & Intelligent Systems*, 11, 407. <https://doi.org/10.1007/s40747-025-02019-z>
- [2] Lee, G., Kim, S., Lee, D., Goh, K., & Kim, H. (2026). Towards efficient language giants: A comprehensive survey on structural optimizations and compression techniques for large language models. *Neural Networks*, 201, 108900. <https://doi.org/10.1016/j.neunet.2026.108900>
- [3] Kim, G. I., Hwang, S., & Jang, B. (2025). Efficient compressing and tuning methods for large language models: A systematic literature review. *ACM Computing Surveys*, 57(10), Article 253, 1–39. <https://doi.org/10.1145/3728636>
- [4] Alozie, E., Imoize, A. L., Olagunju, H. I., Faruk, N., Garba, S., & Olatunji, A. P. (2026). Tiny machine learning for enhanced edge intelligence. In A. L. Imoize, D.-T. Do, & H. Song (Eds.), *Tiny Machine Learning*. Wiley. <https://doi.org/10.1002/9781394294572.ch9>
- [5] El Mrabet, A., Tber, A., Elmousaid, R., Hlou, L., & El Gouri, R. (2025). Tiny machine learning for IoT-enabled embedded systems: A review. In H. Hagrass, Y. Bennani, & M. Nemiche (Eds.), *Intelligent Systems and Advanced Computing Sciences (ISACS 2023)*. Communications in Computer and Information Science, Vol. 2255. Springer. https://doi.org/10.1007/978-3-031-93448-3_6
- [6] Mukweho, H., Mulaudzi, U., Motala, B., & Moyo, S. (2025). Machine learning on microcontrollers for biological sensing: A systematic review [Preprint]. *Research Square*. <https://doi.org/10.21203/rs.3.rs-6844035/v1>
- [7] Gladys, A. A., Aishwarya, R., Vetrivel, V., et al. (2026). Building expert small models: A comprehensive survey of model compression, knowledge distillation, and augmented inference. *Archives of Computational Methods in Engineering*. <https://doi.org/10.1007/s11831-026-10598-4>
- [8] Jiang, Y., Zhao, P., Zhao, C., & Lin, J. (2025). Towards bandwidth efficient edge-cloud collaborative deep learning with data importance driven compression. *Neurocomputing*, 650, 130835. <https://doi.org/10.1016/j.neucom.2025.130835>