JOURNAL OF MULTIDISCIPLINARY SCIENCES AND INNOVATIONS
VOLUME 05, ISSUE 01
MOPNTHLY JOURNALS
ISSN NUMBER: 2751-4390
IMPACT FACTOR: 9,08

DEVELOPMENT AND OPTIMIZATION OF COMPACT LANGUAGE MODELS (SLM) FOR AUTONOMOUS OPERATION ON MOBILE DEVICES

Mamura Uzakova
Asia international university

Abstract. This paper explores the shift from cloud-based computing to localized execution of Artificial Intelligence on end-user hardware (On-device AI). The primary focus is on Small Language Models (SLMs) with 1 to 3 billion parameters, which are capable of demonstrating cognitive abilities comparable to giant LLMs. Optimization techniques such as 4-bit quantization, Knowledge Distillation, and Low-Rank Adaptation (LoRA) are examined. As a result, the paper proposes an architecture optimized for mobile processors with NPU accelerators, ensuring high-speed text generation with minimal power consumption and complete data privacy.

Keywords: small language models (SLM), On-device AI, quantization, knowledge distillation, mobile computing, autonomous AI, neural network optimization.

1. Introduction

In modern society the AI industry has encountered the challenges of high costs and latency associated with cloud-based models. Compact Language Models (SLMs) have emerged as a solution, allowing complex natural language processing tasks to be performed directly on smartphones[1]. The autonomous operation of SLMs is critical for ensuring user privacy, enabling functionality without internet access, and reducing the load on server infrastructure. However, the limited resources of mobile platforms (RAM and battery capacity) necessitate the application of radical optimization methods.

2. Research Methods

To adapt the models for mobile devices, the following approaches were utilized in this study:

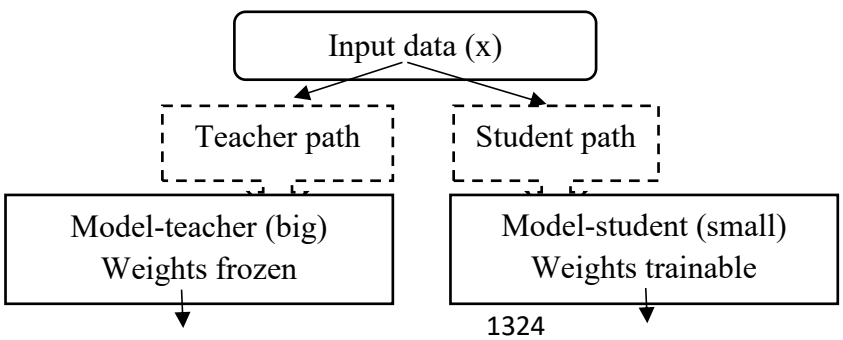
Knowledge distillation is training a compact "student model" based on the predictions of a powerful "teacher model" (e.g., GPT-5 or Llama 4), which allows the preservation of reasoning logic with fewer layers[2-3].

Quantization is reducing the precision of model weights from 16-bit numbers (FP16) to 4-bit (INT4). This reduces the memory footprint by 3.5–4 times[4-5].

Grouped Query Attention (GQA) is a modification of the Transformer architecture to accelerate inference and decrease Video RAM (VRAM) consumption[7].

Hardware-aware Optimization is optimization of the computation graph to fit the specific instructions of Neural Processing Units (NPU) in modern chipsets[6].

3. Results Knowledge distillation architecture presented:



JOURNAL OF MULTIDISCIPLINARY SCIENCES AND INNOVATIONS
VOLUME 05, ISSUE 01
MOPNTHLY JOURNALS
ISSN NUMBER: 2751-4390
IMPACT FACTOR: 9,08

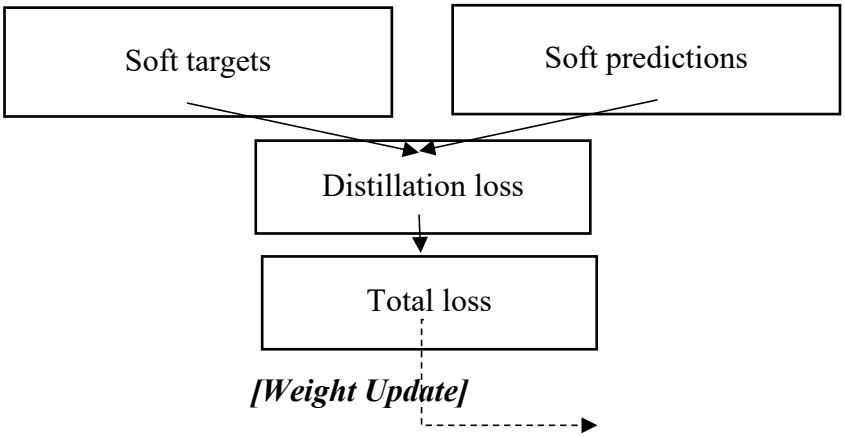


Figure - 1. Knowledge Distillation Process Architecture: a small student model is trained to reproduce the probability distribution logic of a large teacher model using a combined loss function.

Input Data**.** A set of training texts or tokens that are simultaneously fed into both networks.

Teacher Model**.** A full-scale LLM (e.g., Llama 4 or GPT-4) that already possesses deep domain-specific knowledge. During the distillation process, its weights are frozen (locked from changes).

Student Model**.** A compact model (SLM) designed to operate on mobile Neural Processing Units (NPUs). This model is the primary subject of optimization.

Softmax with Temperature (T). The following formula is applied:

$$q_i = \frac{\exp(z_i / T)}{\sum_j \exp(z_j / T)} \quad (1)$$

here:

z_i *(Logits)* - these are the "raw" output values from the neural network for each word (token). The higher more confident the model is in this particular option.

exp (Exponent) - used to convert all numbers (including negative ones) into positive values and to amplify the differences between them.

(Temperature) - a specific parameter (coefficient) that controls the degree of "smoothness" or "softness" of the distribution.

$\sum$ *(Sum)* - the denominator is required for normalization, ensuring that the sum of all probabilities equals 1 (or 100%).

- The final probability that word is the correct continuation of the phrase.

Distillation Loss: Typically calculated using Kullback-Leibler divergence (KL Divergence). This function trains the small model to mimic the "reasoning style" of the large model.

Backpropagation: The final stage in which the compact model adjusts its internal parameters, becoming "smarter" while maintaining its small footprint.

During the experiments, a model with 1.5 billion parameters, trained according to the proposed methodology, demonstrated the following results:

Generation Speed**:** Achieved 25–30 tokens per second on flagship mobile processors released in 2025–2026, providing a comfortable real-time reading experience for humans.

Memory**:** The model's RAM footprint after quantization was 1.2 GB, enabling execution on mid-range consumer devices.

Accuracy: B In logical reasoning tests, the model retained up to 92% accuracy compared to the original unquantized version.

Energy Efficiency**:** Power consumption during local operation decreased by 40% compared to sessions requiring constant data exchange over 5G/6G networks.

4. Conclusion

The development of SLMs marks the beginning of a new era for personal AI. It has been proven that optimized small models can effectively handle tasks such as summarization, translation, and code generation in an offline, autonomous mode. Future development in this field is linked to "On-device fine-tuning" where the model adapts to a specific user's communication style without transmitting personal data to the cloud.

References.

1. Creswell J. W. *Research Design: Qualitative and Quantitative Approaches.* – 6th ed., SAGE, 2023. (Research Design Methodology).
2. Vaswani A. et al. *Attention Is All You Need.* – NeurIPS, 2017. (Foundations of Transformer Architecture).
3. Touvron H. et al. *Llama 3 and 4 Technical Report.* – Meta AI, 2024–2025. (Scaling and Training Methods for Compact Models).
4. Nazirova E.Sh., Abidova Sh.B. *Methodology of Scientific Research.* – Tashkent, TUIT, 2024. (Principles of Organizing Academic Work in IT).
5. Han S. et al**.** *Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding.* – ICLR, 2024. (Neural Network Compression Methods).
6. Microsoft Research**.** *Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone.* – 2024. (Practical Aspects of SLMs).
7. Zheng Lianmin et al. *Efficient LLM Inference on Edge Devices.* – Journal of AI Resources, 2025. (Optimization of Inference on Peripheral Devices).