Adaptive Layer Compression and Storage with QoS-Aware Loading for LLM Serving

Meriem Bouzouad, Vincent Lannurien, Yuan-Hao Chang, Jalil Boukhobza

Adaptive Layer Compression and Storage with QoS-Aware Loading for LLM Serving

Meriem Bouzouad,° Vincent Lannurien,* Yuan-Hao Chang,* Jalil Boukhobza *

* Lab-STICC, CNRS, UMR 6285, ENSTA, Institut Polytechnique de Paris, Brest
° Institute of Information Science, Academia Sinica, Taipei, Taiwan

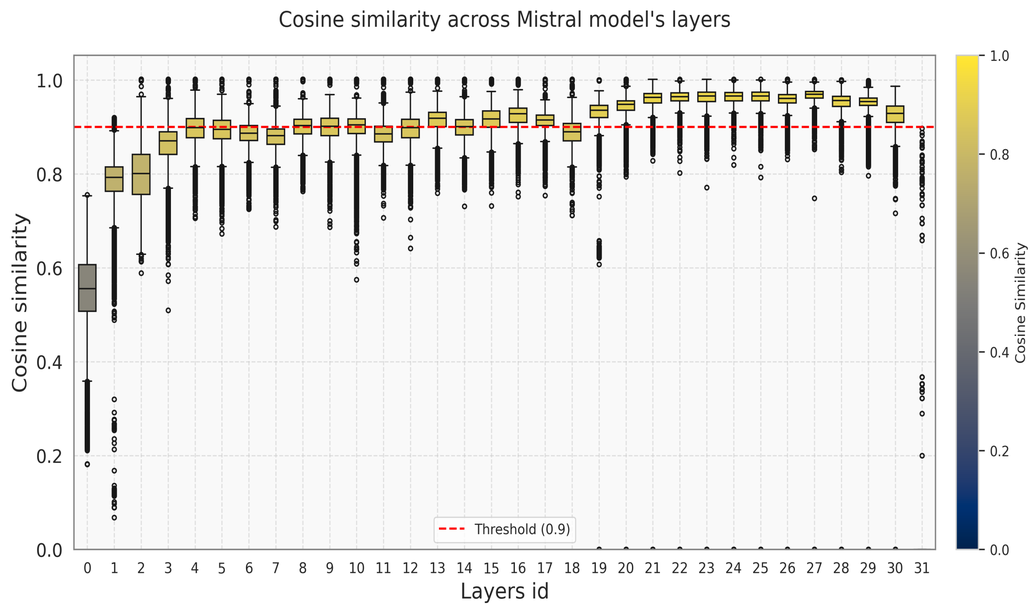
1 – LLM at the Edge

- Pushing generative AI inference to the edge is emerging as a promising approach to enhance privacy and reduce latency [3];
- Although edge accelerators significantly boosted on-device processing capabilities, performance is still constrained by memory;
- Existing methods like quantization, pruning, and low-rank factorization address the issue of LLMs' large memory footprint [5, 1];
- Edge systems have fluctuating workload and different process priorities, which require adapted and intelligent ressource management;
- LLM performance characterization involves a complex interplay of energy usage, inference latency, and accuracy, making unified evaluation a non-trivial task [4].

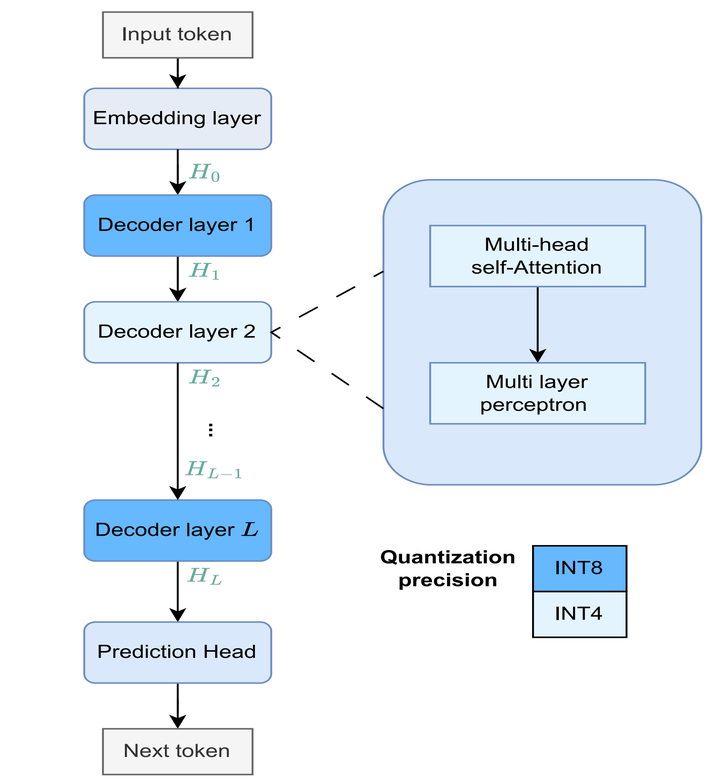
💡 ***Problem Statement*** – **How can we design an edge-optimized solution that accelerates LLM inference to meet real-time workload and Quality of Service (QoS) constraints?**

2 – Motivation

- LLMs are composed of multiple layers; not all of them contribute equally to the output quality [2];
- Cosine similarity is used to measure the distance between the intermediate results (*hidden states*) passed through the layers;
- This analysis assesses whether the transformations applied to hidden states by each layer enhance semantic content or unnecessarily increase computation.

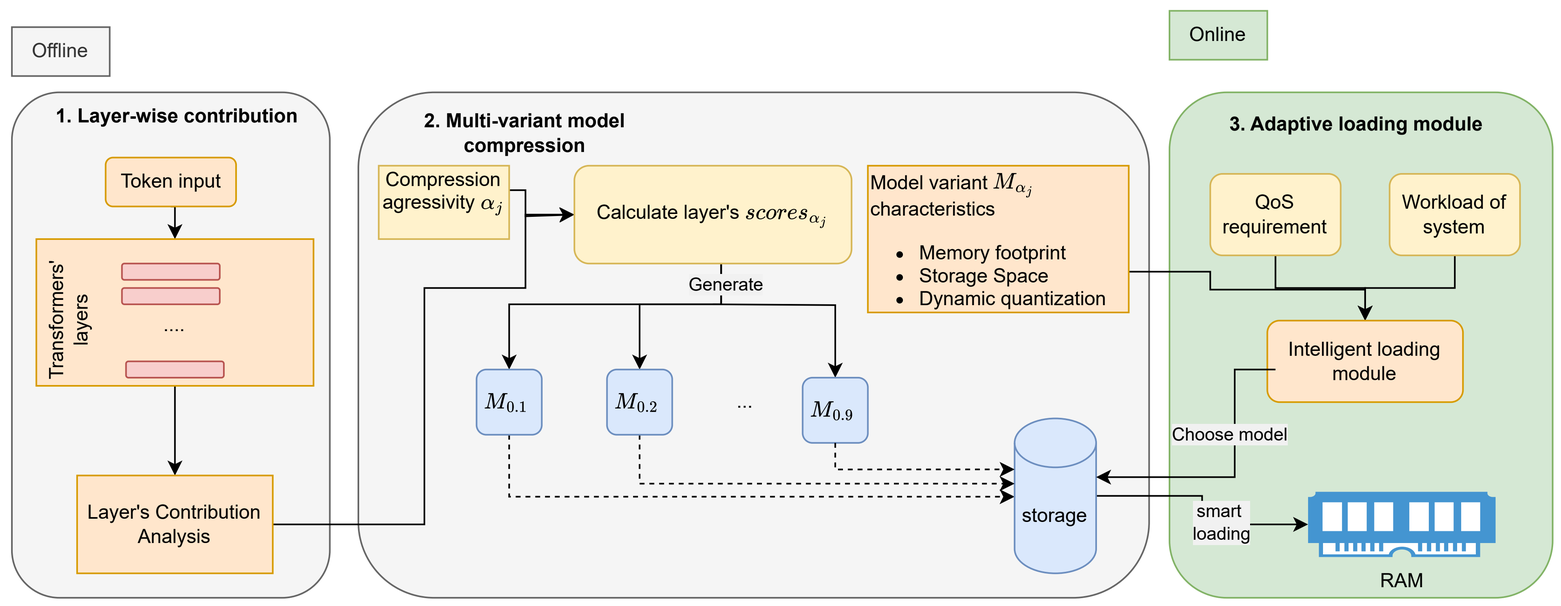


3 – LLM Architecture



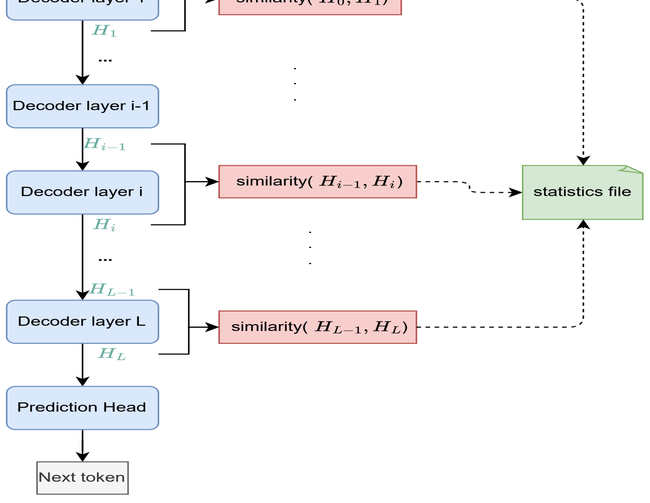
4 – Contribution Overview

❶ Tailoring compression to the importance of layers;
❷ Storing multiple variants depending on the aggressivity of compression with smart storage;
❸ Switching between model variants at runtime to satisfy QoS requirements and dynamic workload changes.



5 – Layer-Wise Contribution

Gathering statistics on data from different tasks allows computing the contribution of each layer, thus enabling layer-wise quantization.



6 – Multi-Variant Model

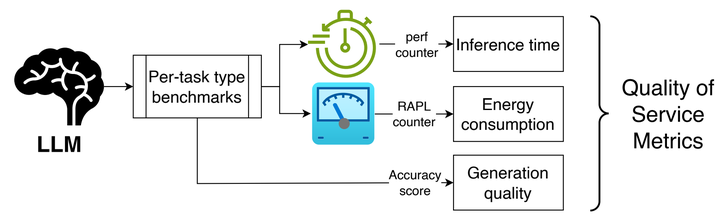
α	0.1	0.5	0.9
Layer 1	INT8	INT8	INT8
Layer 6	INT4	INT8	INT8
Layer 20	INT2	INT3	INT6

Table 1: Different quantization values based on layer scores and compression aggressivity factor α.

- Different model variants exhibit varying performance in terms of inference latency and generation quality;
- Depending on their contribution, some layers maintain the same quantization (INT8, INT5, etc.) across various aggressivity factors, while others may be altered;
- It is essential to provide a storage-aware policy that enables efficient switching between models, by identifying and reusing their shared layers:

$$Storage(Joint{-}Model{-}variants) < \sum_{i}^{n} Storage(M_{\alpha_i})$$

7 – Adaptive Loading



Intelligent loading requires characterization data regarding heterogeneous hardware platforms and models:

- State-of-the-Art benchmarks target various task types: translation, classification, code generation, multi-step problem solving, etc. These can be used to evaluate the generation quality of an LLM;
- Hardware manufacturers provide APIs to measure CPU, DRAM and GPU energy consumption (Intel RAPL, NVIDIA NVML, etc.);
- The OS can help profiling processes to gather performance metrics; instrumenting the code can achieve fine-grained measurements (of *e.g.* the actual inference function call).

[1] Razvan-Gabriel Dumitru et al. "Layer-Wise Quantization: A Pragmatic and Effective Method for Quantizing LLMs Beyond Integer Bit-Levels". In: *arXiv preprint arXiv:2406.17415* (2024).
[2] Shwai He et al. "What Matters in Transformers? Not All Attention is Needed". In: *arXiv preprint arXiv:2406.15786* (2024).
[3] Zhenyan Lu et al. "Small Language Models: Survey, Measurements, and Insights". In: *arXiv preprint arXiv:2409.15790* (2024).
[4] Sasha Luccioni, Yacine Jernite, and Emma Strubell. "Power Hungry Processing: Watts Driving the Cost of AI Deployment?" In: *ACM FAccT*. 2024.
[5] Zhihang Yuan et al. "LLM Inference Unveiled: Survey and Roofline Model Insights". In: *arXiv preprint arXiv:2402.16363* (2024).