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

Meriem Bouzouad[1], Vincent Lannurien[1], Yuan-Hao Chang[2], Jalil Boukhobza[1]

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

Presentation Outline

01

Context

02

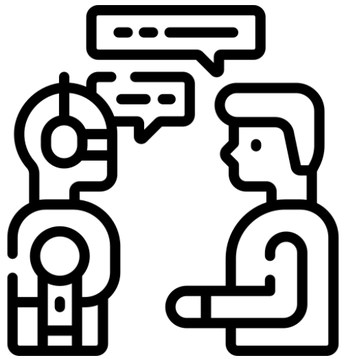
Problem justification

03

Contribution

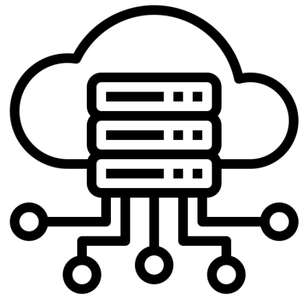
Context

Generative Ai has infiltrated the world due to its ability to approximate or even surpass human intelligence.

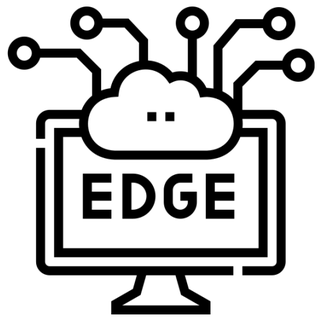


“Who doesn’t use generative AI? Exactly.”

- Latency
- Privacy



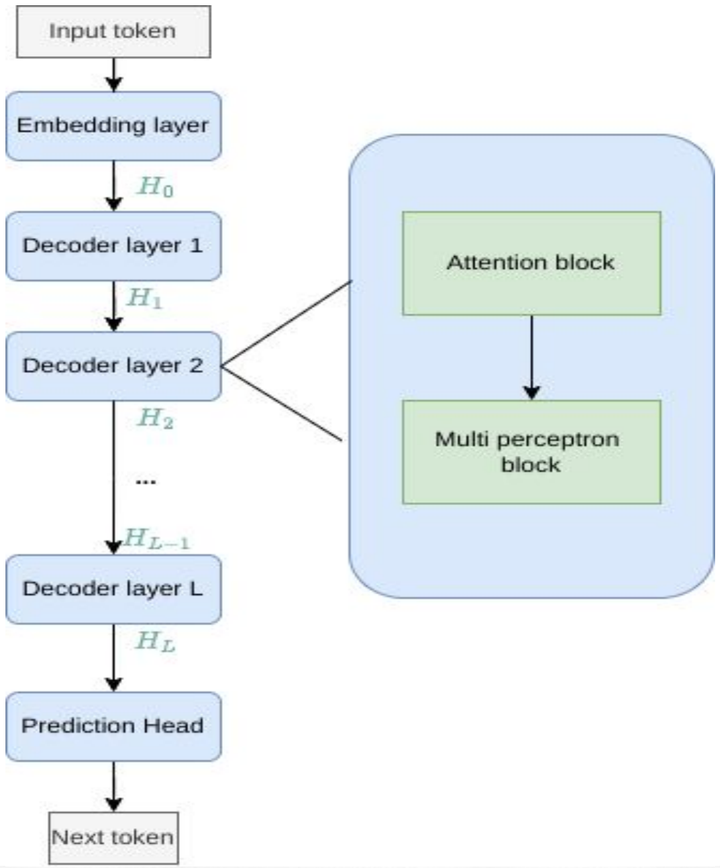
Gen AI is being pushed towards Edge [1]



- High computing demands
- Memory constrained
- Smaller storage space

Problem justification

“Behind the success of generative AI”



- They are Large: they consume memory, compute and storage.
- Optimizations include: Quantization, pruning, distillation..etc [2-3]

⇒Tradeoff between latency and generation quality [2].

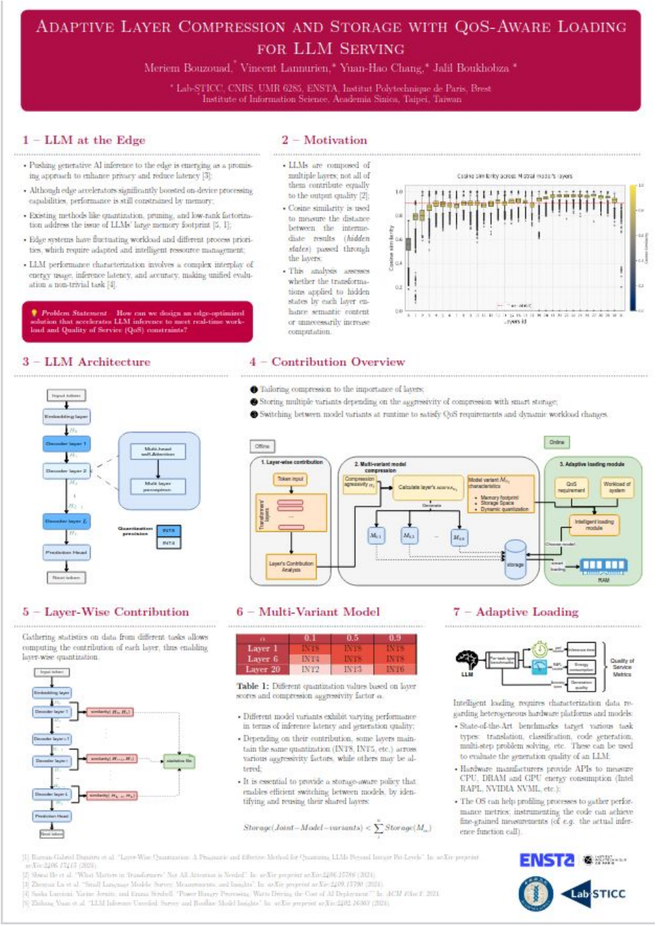
Rule of thumb: Take a model that can fit in memory. However edge workload vary and static model deployment doesnt can underestimate or overestimate system’s needs.

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?

Contribution

- Dynamic layer compression depending on its importance.
- Leverages a pool of models with different sizes to adapt to QoS dynamically such as latency, energy, and accuracy.
- Efficient storage to switch between models on demand.

Intelligent compression + Adaptability.



References

[1] Zhenyan Lu et al. “Small Language Models: Survey, Measurements, and Insights”. In: arXiv preprint arXiv:2409.15790 (2024).
[2] Zhihang Yuan et al. “LLM Inference Unveiled: Survey and Roofline Model Insights”. In: arXiv preprint arXiv:2402.16363 (2024)
[3] 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).

Thank you