LLM Size Reduction and Carbon Footprint

Pierre Dosquet, Ashwin Ittoo
HEC - University of Liège, Belgium

Abstract

Compression techniques like quantization reduce memory and speed up inference for LLMs, but their environmental impact during inference is underexplored. This study quantifies how 4-bit quantization affects performance and CO_2-equivalent emissions across hardware and electricity mixes using LLaMA-7B/30B and Mistral-7B-v0.3/Small 3.

Introduction

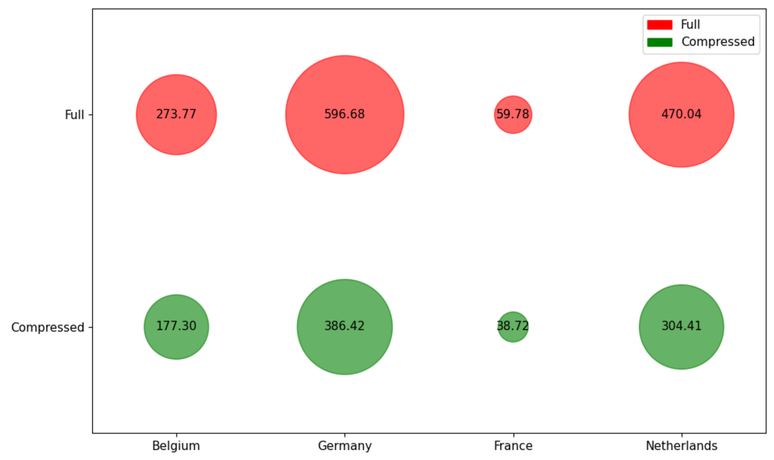
While LLMs excel in NLP tasks, their inference energy and emissions remain understudied. We present the first empirical study linking quantization to performance and environmental impact.

Methodology

- **Models:**
  - LLaMA-7B/30B
  - Mistral-7B-v0.3/Small 3
- **Techniques:** 4-bit OPTQ quantization
- **Metrics:** WikiText-2, MMLU, IFEval
- **Setups:**
  - Setup 1: AMD EPYC 7513, NVIDIA A100 SXM4 80GB, 240GB RAM
  - Setup 2: AMD EPYC 7513, NVIDIA A100 40GB, 60GB RAM
- **Energy Measurement:** CodeCarbon
- **Emissions Calculation:** Regional grid intensities

Results

- **Negligible Accuracy Loss:** 4-bit quantization of LLaMA and Mistral models shows minimal impact on task performance.
- **Hardware-Dependent Energy Effects:** Energy consumption varies from a 39% decrease to a 26% increase depending on hardware setup.
- **Geographic Dependence:** Compressed models in carbon-intensive grids can emit up to 6 times more CO_2 than uncompressed models in low-carbon grids.



Conclusion

While 4-bit quantization preserves performance, sustainability benefits depend on hardware and grid carbon intensity.

- Quantization can reduce or increase energy consumption depending on hardware.
- Geographic location significantly impacts CO_2 emissions.

References

[1] CodeCarbon (2024)
Electricity Maps (2025)
OPTQ: Accurate Quantization for Generative Pre-trained Transformers (2023)