LLM Size Reduction and Carbon Footprint

Pierre Dosquet[1 [0009−0004−0853−7264]] and Ittoo Ashwin[1 [0000−0001−9069−5018]]

HEC – University of Liège, Belgium
pdosquet@student.uliege.be, ashwin.ittoo@uliege.be

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. Results show negligible accuracy loss but hardware-dependent energy effects (−39% to +26%) and strong geographic dependence: compressed models in carbon-intensive grids can emit up to 6× more CO_2eq than uncompressed models in low-carbon grids. These findings link compression, hardware efficiency, and grid context, advocating for carbon-aware LLM deployment.

Keywords: LLM compression · quantization · energy efficiency · carbon footprint

1 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.

2 Methodology

We evaluate LLaMA-7B/30B and Mistral-7B-v0.3/Small 3 at full precision and with 4-bit OPTQ quantization [3]. Performance is assessed via WikiText-2, MMLU, and IFEval. Energy consumption (CPU/GPU/RAM) is measured using CodeCarbon [1] on two setups: Setup 1 (AMD EPYC 7513 32-core with 2 active cores, NVIDIA A100 SXM4 80GB, 240GB RAM) and Setup 2 (AMD EPYC 7513 32-core with 2 active cores, NVIDIA A100 40GB, 60GB RAM). CO_2eq emissions were computed using regional grid intensities [2].

3 Results

Results show negligible accuracy loss post-quantization but hardware-dependent energy effects (−39% to +26%). Emissions vary strongly by geography: compressed models in carbon-intensive grids emit up to 6× more CO_2 than uncompressed models in low-carbon grids.

Model	Setup	CPU	GPU	RAM	Hardware
LLaMA-7B	1	+65%	-18%	0%	+26%
	2	-40%	-28%	0%	-28%
Mistral-7B-v0.3	1	+57%	-18%	0%	+22%
	2	-46%	-27%	0%	-31%
LLaMA-30B	1	+58%	-41%	0%	+1%
	2	-42%	-48%	0%	-39%
Mistral Small 3	1	+58%	-33%	0%	+8%
	2	-42%	-42%	0%	-35%

Table 1. Percentage variation in energy consumption of quantized models (4-bit) relative to the full-precision baseline (Setup 1). Positive values indicate increased consumption; negative values indicate savings.

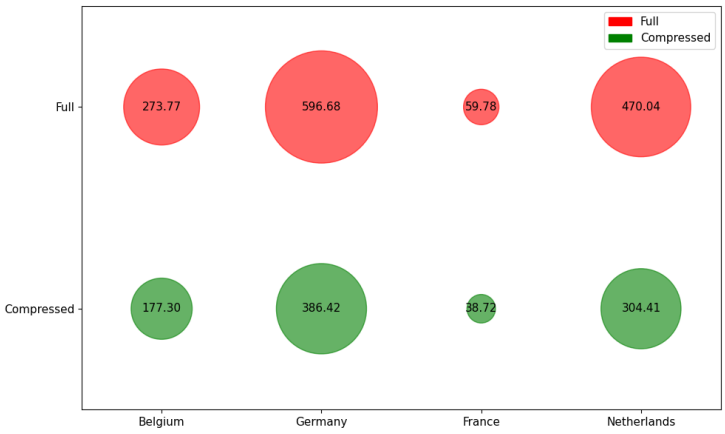

Fig. 1. CO_2eq emissions (gCO_2e) for full precision Mistral Small 3 model (Setup 1) vs. compressed model (Setup 2) across regions.

4 Conclusion and Future Work

This work links LLM compression to energy efficiency and geography. While 4-bit quantization preserves performance, sustainability benefits depend on hardware and grid carbon intensity. Key contributions include: (i) empirical evaluation of quantization's energy/emission effects; (ii) a grid-aware methodology; and (iii) guidance for sustainable LLM deployment. Future work will explore compression methods and more realistic inference workloads.

References

1. Courty, B., Schmidt, V., Luccioni, S., Goyal-Kamal, MarionCoutarel, Feld, B., Lecourt, J., LiamConnell, Saboni, A., Inimaz, supatomic, Léval, M., Blanche, L., Cruveiller, A., ouminasara, Zhao, F., Joshi, A., Bogroff, A., de Lavoreille, H., Laskaris, N., Abati, E., Blank, D., Wang, Z., Catovic, A., Alencon, M., Michał Stęchły, Bauer, C., de Araújo, L.O.N., JPW, MinervaBooks: mlco2/codecarbon: v2.4.1 (May 2024). https://doi.org/10.5281/zenodo.11171501, https://doi.org/10.5281/zenodo.11171501
2. Electricity Maps: Electricity Maps | the world's most comprehensive electricity data platform. https://www.electricitymaps.com/ (2025), accessed: 2025-05-03
3. Frantar, E., Ashkboos, S., Hoefler, T., Alistarh, D.: OPTQ: Accurate quantization for generative pre-trained transformers. In: The Eleventh International Conference on Learning Representations (2023), https://openreview.net/forum?id=tcbBPnfwxS