

Efficient LLMs Inference for Green AI:

Energy-Efficient Deployment of Large Language Models (LLMs): A Systematic Benchmarking Study for Green AI

Supervisors: Fatima Rani, Pit Hofmann

Context:

Large Language Models (LLMs) have become central to modern AI systems, driving applications in natural language understanding, reasoning, and generation. However, their immense computational and energy demands raise sustainability concerns. The field of Green AI seeks to optimize the trade-off between performance and environmental cost by evaluating efficiency across hardware and algorithmic dimensions.

This thesis investigates the energy efficiency, latency, and accuracy trade-offs of deploying LLMs across heterogeneous hardware platforms (e.g., GPUs, NPUs, edge accelerators). By systematically benchmarking transformer-based architectures under varied inference configurations, the project contributes empirical evidence toward sustainable deployment strategies for LLMs and provides insights into the environmental impact of large-scale language models.



Green AI KPI's & Metrics

Tasks:

- Implement and fine-tune LLMs using efficient frameworks (e.g., Hugging Face Transformers, DeepSpeed, vLLM).
- Benchmark models on representative NLP tasks such as text classification, summarization, and question answering.
- Compare lightweight and full-scale LLM variants (e.g., LLaMA) under controlled conditions.
- Deploy models on diverse hardware backends (GPUs, edge accelerators, and cloud-based inference engines).

- Evaluate optimization strategies for sustainable inference, including quantization, pruning, low-rank adaptation (LoRA), and distillation.
- Analyze multi-objective trade-offs between energy consumption, response latency, and output quality.
- Derive actionable guidelines for energy-efficient deployment of LLMs in real-world environments.

KPIs and Metrics:

- Accuracy & Quality:
 - Task-specific metrics such as F1 score, BLEU, ROUGE, or perplexity.
 - Output coherence and factuality metrics for generative tasks.
- Efficiency & Energy Metrics:
 - Energy per token / per inference (Joules).
 - Throughput (tokens or sequences per second).
 - Latency (average response time per request).
 - Hardware utilization efficiency (% GPU/NPU usage).
 - Energy-Performance Pareto analysis (trade-off curves).
- Sustainability Indicators:
 - Carbon footprint estimation (gCO₂e per inference).
 - Model size vs. efficiency scaling laws.
 - Green AI index: composite metric integrating accuracy, energy, and latency.

Required skills:

- Python programming skills.
- Proficiency with deep learning frameworks (PyTorch, TensorFlow).
- Willing to learn LLM training/inference pipelines & optimization techniques (quantization, distillation, LoRA).
- Basic understanding of sustainability metrics and Green AI principles.

Key words: **Green AI;** Large Language Models (LLMs), Energy Efficiency; Inference Optimization, Sustainable Benchmarking, Green AI.

Language: English

Corresponding email: fatima.rani@tu-dresden.de

References :

1. Jegham, N., Abdelatti, M., Elmoubarki, L., & Hendawi, A. (2025). How hungry is ai? benchmarking energy, water, and carbon footprint of llm inference. arXiv preprint arXiv:2505.09598.
2. Khan, T., Motie, S., Kocak, S. A., & Raza, S. (2025). Optimizing Large Language Models: Metrics, Energy Efficiency, and Case Study Insights. arXiv preprint arXiv:2504.06307.
3. Stojkovic, J., Choukse, E., Zhang, C., Goiri, I., & Torrellas, J. (2024). Towards greener llms: Bringing energy-efficiency to the forefront of llm inference. arXiv preprint arXiv:2403.20306.