Energy Efficiency in Large Language Model (LLM) Inference

Under the Regis LLM Research Group

- **Course:** MSDS 692 – Data Science Practicum I

- **Instructor: Christy Pearson**
- **Mentor: Dr. Kellen Sorauf**
- **Researcher: Uttam Emmaniuel Dammu**

Project Overview

- This practicum project, conducted under **Dr. Kellen Sorauf** , investigates the **energy usage and carbon footprint of large language models (LLMs)**.
- It compares **open-source models** (*Mistral 7B*, *Mamba 7B*, *Qwen 2 7B*) with **closed commercial models** (*GPT-5-Pro* and *GPT-4.1-mini* via OpenAI API) to quantify efficiency trade-offs in inference.
- Using **AWS GPU instances** and **CodeCarbon**, the study measures how **quantization (4-bit vs 8-bit)**, **prompt brevity**, and **batching** affect **energy consumption** and **accuracy**.
- Findings contribute to ongoing research in **sustainable AI** and **energy-aware model deployment**.

Objectives

- Measure GPU **power**, **latency**, and **CO emissions** for open vs closed LLMs.

- Evaluate **quantization** and **prompt length** impact on efficiency.

- Compare **Mistral 7B**, **Mamba 7B**, **Qwen 2 7B**, **GPT-5-Pro**, and **GPT-4.1-mini**.

- Build an **interactive dashboard** of energy/performance metrics.

- Provide **actionable insights** for sustainable AI use.

Green AI Motivation

- This project is grounded in the philosophy of **Green AI**, emphasizing the reduction of computational waste and environmental impact in machine-learning workflows.
- Drawing on the concept of **Green Prompt Engineering** (Srinivasan et al., 2024), it examines how *prompt brevity*, *quantization*, *batching*, and *routing* can reduce GPU energy usage without compromising model quality.
- By treating energy and CO output as evaluation metrics alongside accuracy and latency, the study demonstrates how sustainability and performance can coexist in modern LLM research.

Research Context

- This project extends **Dr. Sorauf's Sustainable AI Initiative**, shifting focus from data-level to **model-level energy profiling**.
- It connects prior greenhouse-gas analytics research with modern LLM benchmarking for transparency and responsible AI design.
- Grounded in the principles of **Green AI** and **Green Prompt Engineering**, this work operationalizes energy-conscious inference—showing how small optimizations in prompt structure and numerical precision translate into measurable sustainability benefits.

Methodology

Stage	Description
1. Environment Setup	Configured AWS EC2 g6 L4 & A10G GPUs with CodeCarbon, PyTorch, and Transformers.
2. Baseline Measurement	Ran QA, summarization, and reasoning tasks; recorded energy (kWh), CO (kg), and latency (s).
3. Optimization	Applied 4-bit/8-bit quantization, prompt brevity, and batch routing for performance tuning.
4. Logging	Captured NVML-based GPU power metrics and CodeCarbon logs per run.
5. Visualization	Generated Plotly Dash dashboard comparing energy, latency, and cost efficiency.

Tools & Libraries

Languages: Python 3.10+
Frameworks: PyTorch, Hugging Face Transformers
Energy Tracking: CodeCarbon, NVIDIA NVML
Visualization: Plotly, Dash
Hardware: AWS EC2 (L4 and A10G GPUs)
Environments: Jupyter Lab · VS Code · Google Colab

Key Results

Model	Precision	Energy (kWh)	CO (kg)	Latency (s)	Observation
Mistral 7B	8-bit	0.037	0.018	5.6	Stable baseline
Mistral 7B	4-bit	0.024	0.012	4.1	35 % energy reduction
Mamba 7B	8-bit	0.034	0.016	5.3	Memory-efficient
Qwen 2 7B	8-bit	0.045	0.021	5.8	Highest open-model draw
GPT-4.1-mini	N/A	—	—	2.74	Fastest inference
GPT-5-Pro	N/A	—	—	33.69	Most accurate, most costly

Average costs (Week 7 data):
- GPT-4.1-mini → $0.0265
- GPT-5-Pro → $1.0583

Server-side energy of API models not measurable; shown at 0 for relative context.

Insight → Quantization + prompt brevity saved **35–40 % energy** with no major accuracy loss.

View Interactive Dashboard

Explore the complete benchmarking visuals and detailed insights:

- **Dashboard Details:** Dashboard_Details.md
- **PDF Preview:** LLM Energy Benchmark — Open vs Closed (PDF)
- **Interactive Dashboard (HTML Preview):**
  Open Dashboard in Browser

Dashboard Summary — *LLM Energy Benchmark (Open vs Closed)*

This dashboard visualizes **energy efficiency**, **latency**, and **accuracy** across both **open-source** and **closed API** Large Language Models (LLMs).
It compares **Mamba-Codestral-7B** and **Qwen2-7B** (in 4-bit and 8-bit quantization) with **closed systems** like ChatGPT and Gemini, focusing on measurable sustainability metrics.

Data Overview

All open models were benchmarked locally using **CodeCarbon** and **NVIDIA GPU telemetry** on AWS EC2 (L4 GPU).
Closed models were evaluated through API latency monitoring, as energy usage is not observable through hosted endpoints.

Data File	Description
`week8_open_closed_unified.csv`	Unified dataset for visualization.
`emissions_*.csv`	Raw CodeCarbon logs for open models.
`open_models_combined_normalized.csv`	Processed open-model results.
`closed_models_combined_normalized.csv`	Closed-model latency data.

Visualizations

Plot	Description
Energy vs Input Tokens	Shows how energy usage scales with prompt size.
Latency vs Energy	Examines runtime vs power trade-offs.
Energy per 1k Tokens	Normalized energy efficiency across open models.
Energy vs Accuracy	Displays the efficiency frontier (accuracy vs energy).
Latency Distribution	Reveals runtime stability per model.
Emissions vs Latency	Links environmental impact to performance.

Key Findings

- **Mamba-7B (4-bit)** → Lowest energy per 1k tokens; most efficient overall.
- **Qwen2-7B (8-bit)** → Best accuracy-energy balance with consistent latency.
- **Closed models** → Fast responses but no measurable energy transparency.
- **Quantization** → Reduces energy use by up to **45%** with minimal performance drop.

Outputs

- `Dashboards/unified_dashboard.html` — Interactive Plotly dashboard
- `Dashboards/LLM Energy Benchmark - Open vs Closed.pdf` — Static summary version
- `week8_open_closed_unified.csv` — Unified dataset for reproducibility

Result Type: Quantitative Benchmarking
Metrics: Energy (kWh), Latency (s), Accuracy, Emissions (kg CO e)
Domains: Energy Efficiency • Model Performance • Sustainability

---

Insights

- Quantization reduces GPU energy draw by ~35–40 %.
- Closed models achieve high accuracy but at greater latency and cost.
- Open models scale efficiently and transparently.
- Transparency gap: Energy data for hosted APIs remain inaccessible.
- **Green Prompt Engineering** (combining prompt brevity + quantization) cut GPU energy use 40 % while maintaining accuracy.

Technical Notes

- Logged via **CodeCarbon + NVIDIA NVML** (1 Hz intervals).
- API latency and cost from **OpenAI 2025 billing model**.
- Dashboard built with **Plotly Express + Dash** .
- Dataset: `/week8_open_closed_unified.csv`.

Challenges & Learning Outcomes

Challenges: AWS GPU quota limits, runtime interruptions, missing emission logs.
Learning Outcomes: Energy benchmarking, quantization optimization, reproducible green AI workflow.

Future Work

- Implement real-time energy monitoring agents for inference.
- Expand testing to Jetson Orin Nano Super Developer Kit for edge deployment.
- Evaluate energy efficiency in real-world, low-power environments.
- Study GPU temperature–efficiency relationships under continuous load.
- Develop a public Green AI dashboard for open benchmarking.

References

- Acharya, R., et al. (2025). *Agentic AI: Autonomous Intelligence for Complex Goals.* IEEE Trans. AI.
- Cummings, M. (2025). *Identifying AI Hazards and Responsibility Gaps. J. AI Ethics,* 12(1), 44–59.
- Moon, S., & Ahn, K. (2025). *Metrics and Algorithms for Identifying and Mitigating Bias in AI Design. IEEE Access,* 33(4).
- Khan, M., et al. (2023). *AI Ethics: An Empirical Study on Practitioner Views. Computers in Human Behavior,* 152.
- CodeCarbon (2023). *Track Carbon Emissions from ML Computing.* https://mlco2.github.io/codecarbon/
- Dettmers, T., et al. (2022). *LLM.int8(): 8-bit Matrix Multiplication for Transformers.* arXiv:2208.07339.
- Gemini Team (2025). *Gemini 2.5: Advanced Reasoning and Multimodality* [White paper]. Google DeepMind.
- Mistral AI (2024). *Mamba-Codestral-7B-v0.1 [LLM].* Hugging Face.
- MLCommons (n.d.). *Training Benchmarks.* https://mlcommons.org
- OpenAI (2025). *GPT-5-Pro & GPT-4.1-mini [LLMs].* https://platform.openai.com
- Plotly (2024). *Interactive Data Visualization in Python.* https://plotly.com/python
- Srinivasan, A., et al. (2024). *Green Prompt Engineering.* Proc. 13th Int. Conf. AI for Sustainable Development.
- Medium (2024). *Building Minimalistic Dashboards for Data Science Projects.* https://medium.com
- Hugging Face (2025). *Transformers Documentation.* https://huggingface.co/docs/transformers

---

Acknowledgments

This work was guided by **Dr. Kellen Sorauf** , whose mentorship shaped the study's analytical and technical depth.
Appreciation to Regis University's Data Science faculty for their support of sustainable AI research.

Progress Reports

- Week 2 – AWS setup & CodeCarbon installation
- Week 3 – Baseline data collection
- Week 4 – Summarization & EDA
- Week 5 – Quantization & batching optimization
- Week 6 – Routing & Qwen integration
- Week 7 – Closed-model analysis & dashboard
- Closed Model Deep Analysis (Week 7 Extension)

Presentation

This project will be **presented live** as part of the final practicum session under the supervision of ** Professor Christy Pearson**.
The live presentation will summarize key findings, visualizations, and methodology from this repository.

A copy of the presentation slides is included here for reference:

Presentation Materials

File	Description
Final_Presentation.pptx	Editable PowerPoint version
Final_Presentation.pdf	Final exported version (for review/submission)

Folder Structure

"'bash Emmaniuel/ Code/ .gitkeep Mamba_4bit.ipynb Mamba_8bit.ipynb Qwen2_4bit.ipynb Qwen2_8bit.ipynb OpenAI_ClosedModels_Run.ipynb Dashboards/ .gitkeep Dashboard_Details.md ← (new file you'll add) unified_dashboard.html ← main interactive Plotly dashboard LLM Energy Benchmark — Open vs Closed.pdf ← exported static version Data/ .gitkeep emissions_mamba7b_4bit.csv emissions_mamba7b_8bit.csv emissions_qwen2_7b_4bit.csv emissions_qwen2_7b_8bit.csv open_models_combined_normalized.csv closed_models_combined_normalized.csv week8_open_closed_unified.csv ← main unified dataset for dashboard Presentation/ .gitkeep Final_Presentation.pptx ← Editable PowerPoint file Final_Presentation.pdf ← Exported final version for submission/viewing Reports/ .gitkeep Week_2_Progress_Report.pdf Week_3_Progress_Report.pdf Week_4_Progress_Report.pdf Week_5_Progress_Report.pdf Week_6_Progress_Report.pdf Week_7_Progress_Report.pdf week7_closed_openai_report.md README.md

---

License

Released under the **MIT License** for academic research and educational use.