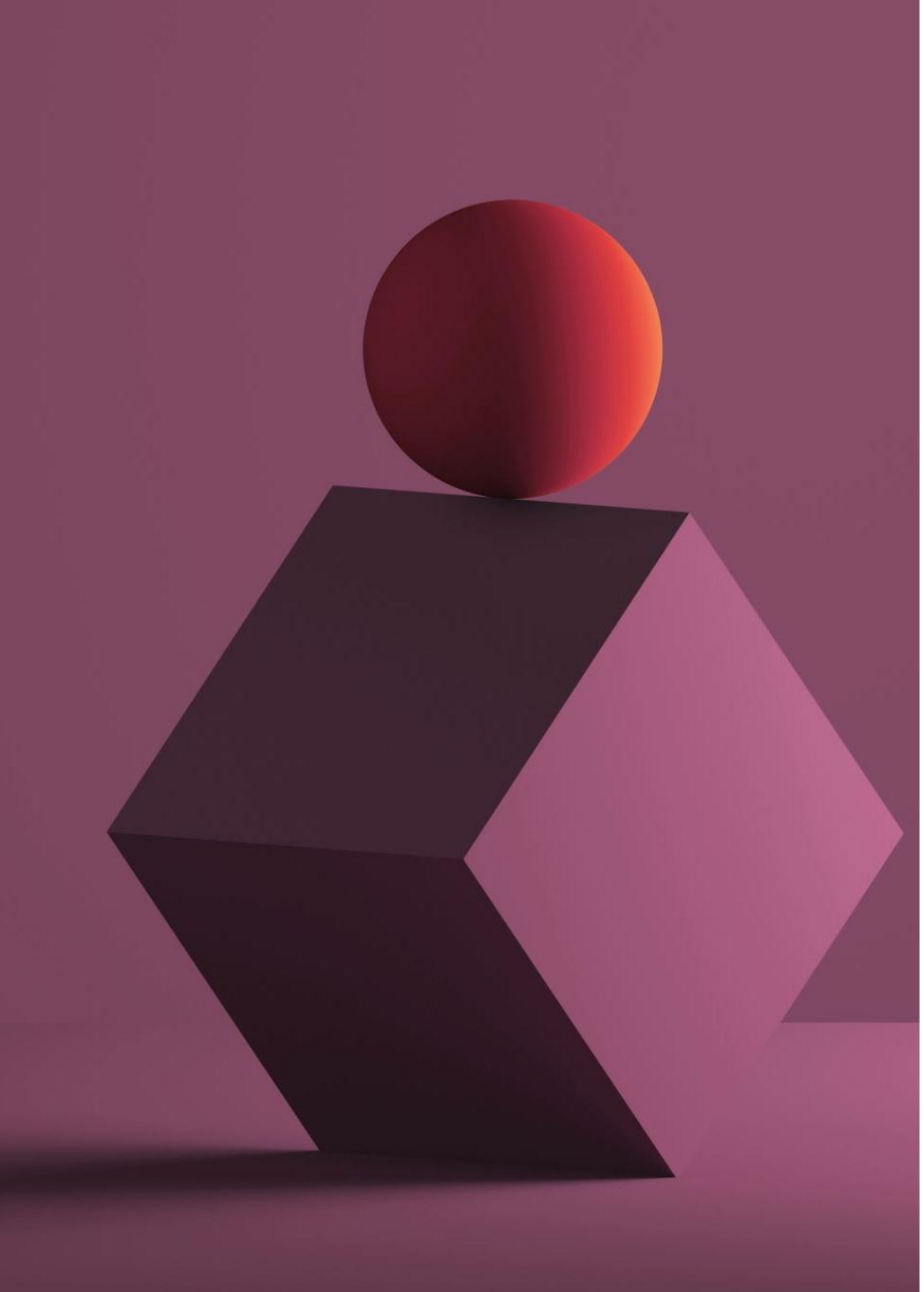


ENERGY EFFICIENCY IN LARGE LANGUAGE MODEL (LLM) INFERENCE: OPEN VS CLOSED MODELS

- *REGIS LLM RESEARCH GROUP – SUSTAINABLE AI BENCHMARKING PROJECT*

Uttam Emmaniuel Dammu

Anderson College of Business and Computing, Regis University

MSDS 692 – Data Science Practicum

Professor: Christy Pearson

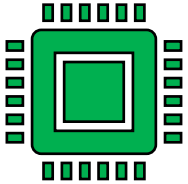
Mentor: Dr. Kellen Sorauf

October 16, 2025

RESEARCH QUESTIONS

How much energy does a large language model consume per inference?

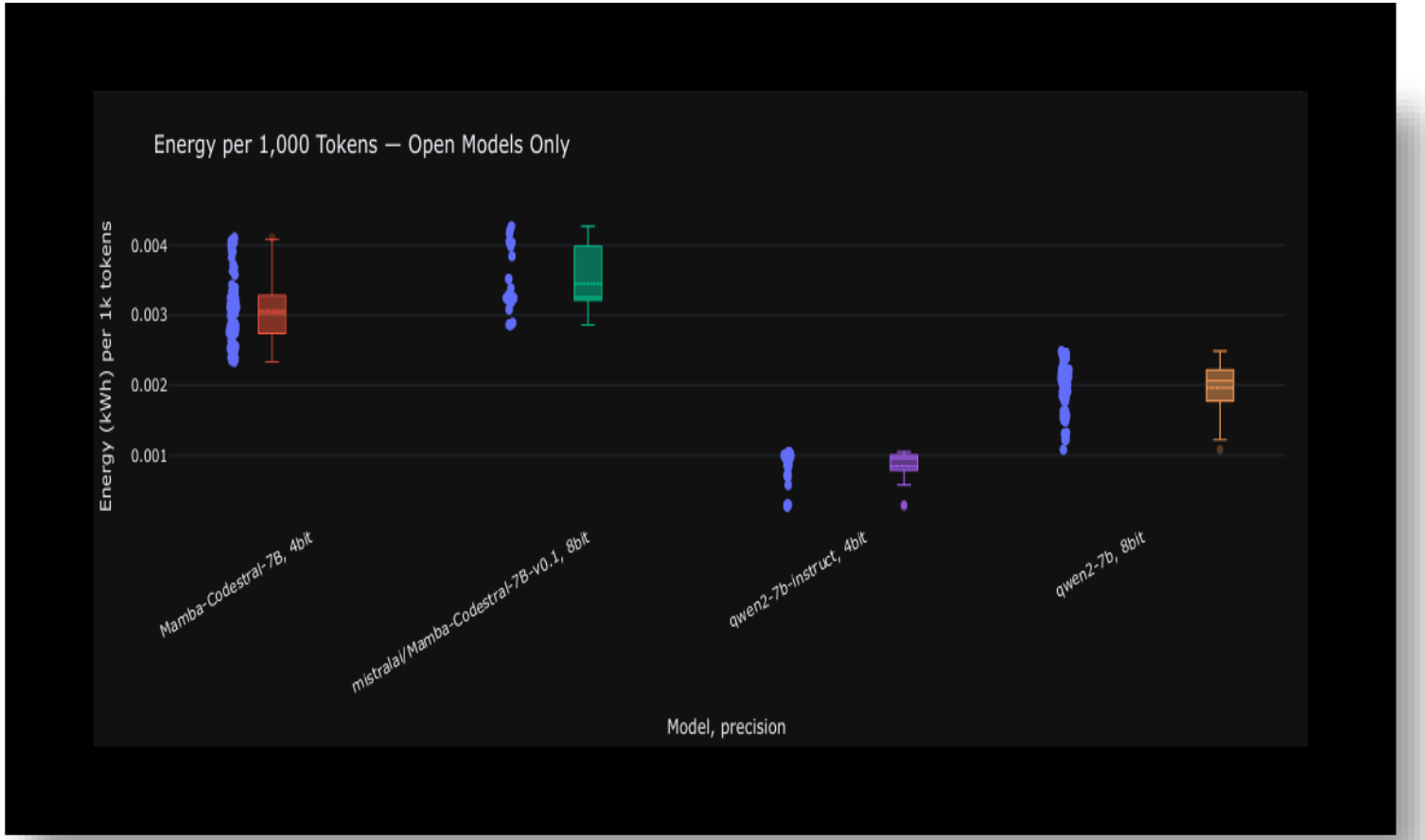
How do model size, prompt length, and quantization affect energy and latency?



Can optimization techniques (prompt brevity, batching, routing) reduce energy use?

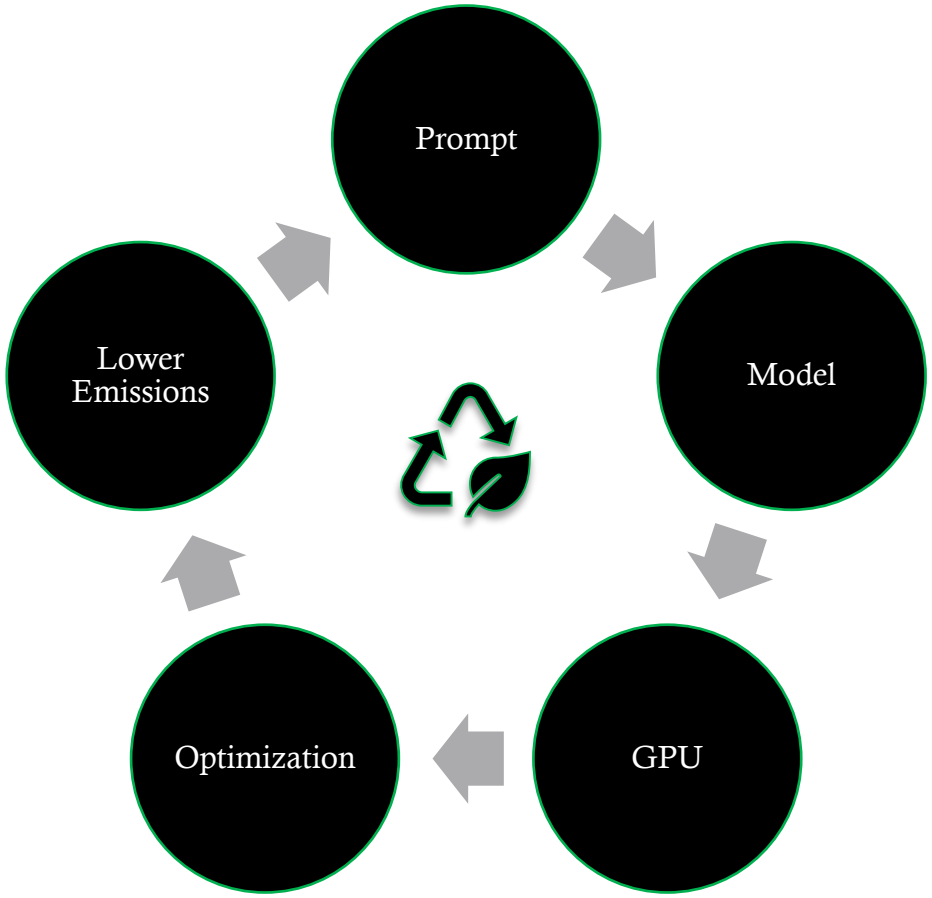
How do open-source and closed models compare in transparency, cost, and performance?

GREEN AI



- Explosive growth in LLM usage → increased energy demand
- AI training + inference = measurable carbon footprint
- Green AI aims to reduce environmental cost of computation
- Limited research on **inference-side** energy consumption

GREEN AI: TOWARDS SUSTAINABLE INTELLIGENCE



AI models use significant energy during every response.

Green AI focuses on reducing that energy use without losing accuracy.

This project improves efficiency through:

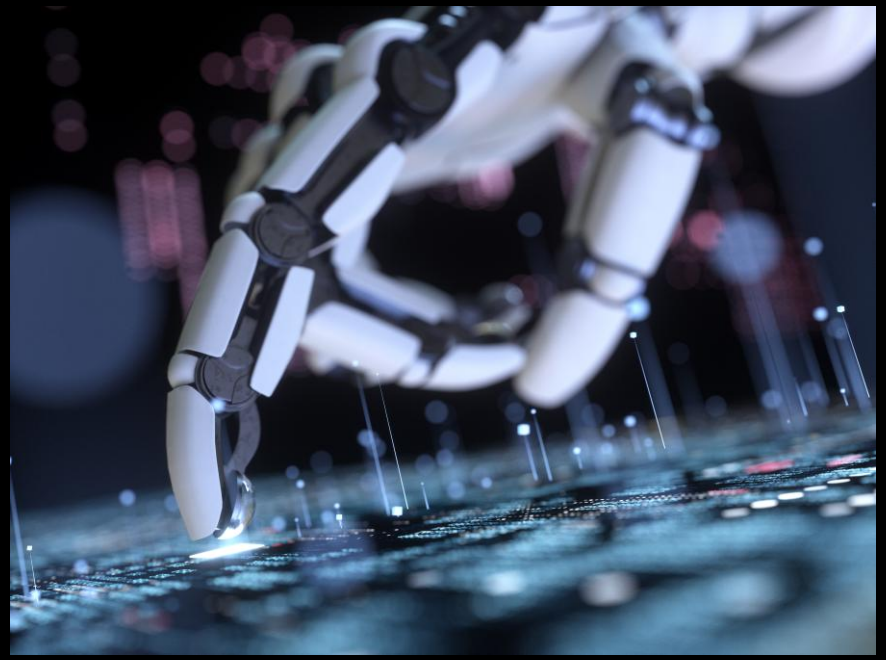
- **Model Precision:** Using lighter (4-bit / 8-bit) versions.

- **Prompt Design:** Writing shorter, focused prompts.

- **Batching:** Grouping similar requests for smoother processing.

Goal: Make AI cleaner, faster, and more energy-conscious.

PROJECT OBJECTIVES



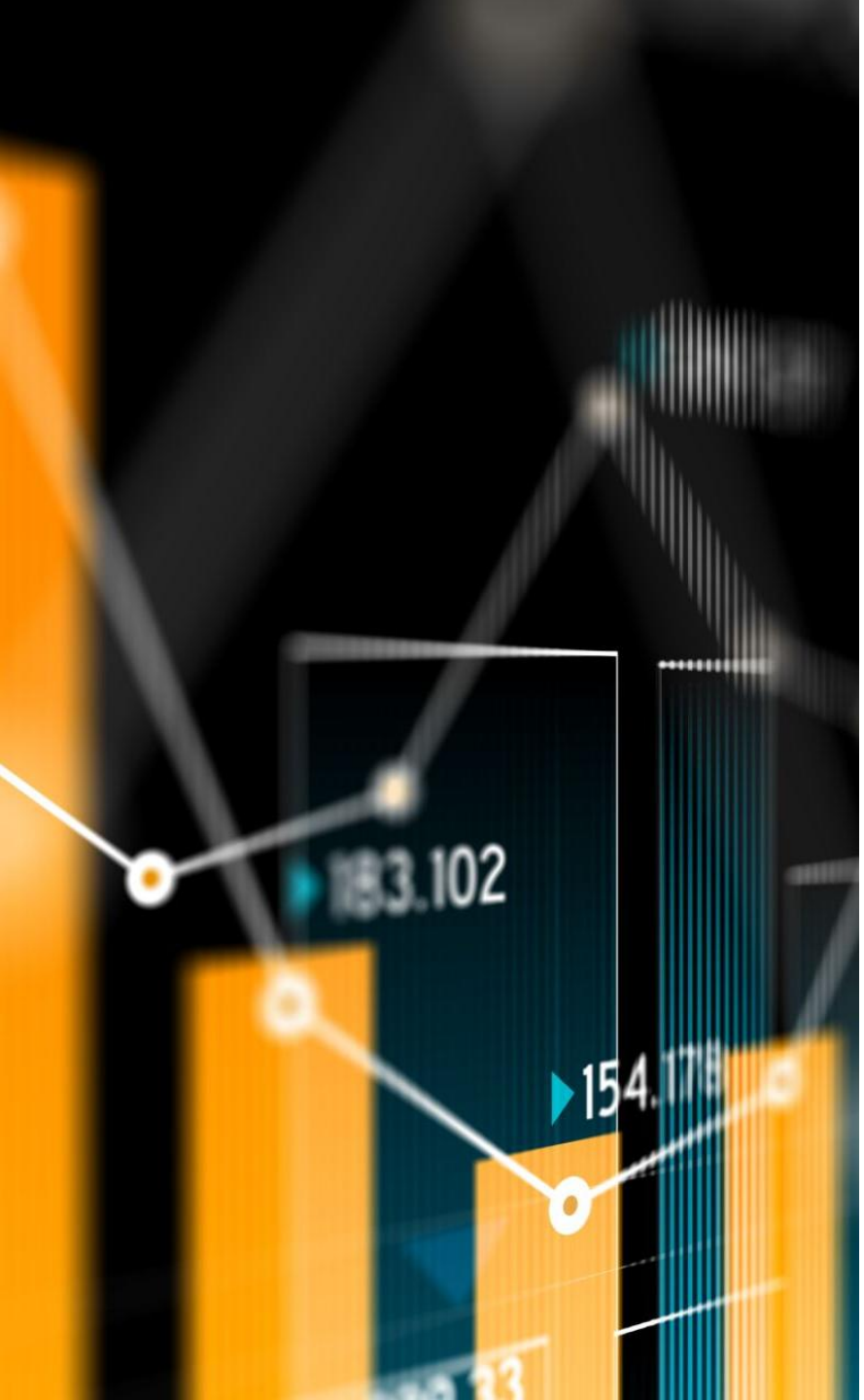
Quantify GPU energy and CO_2 emissions during LLM inference.

Compare open-source (Mistral-Mamba, Qwen2) vs closed (GPT-4.1-mini, GPT-5-Pro).

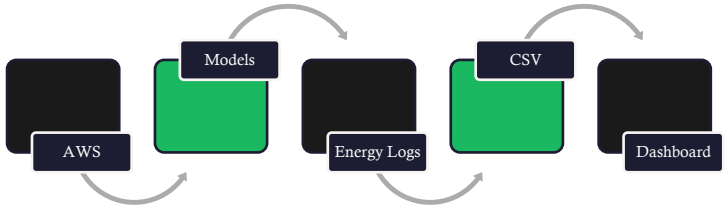
Analyze effects of quantization (4-bit vs 8-bit) and prompt brevity.

Build an interactive dashboard visualizing efficiency trends.

Derive insights for sustainable, “Green” AI deployment.



METHODOLOGY

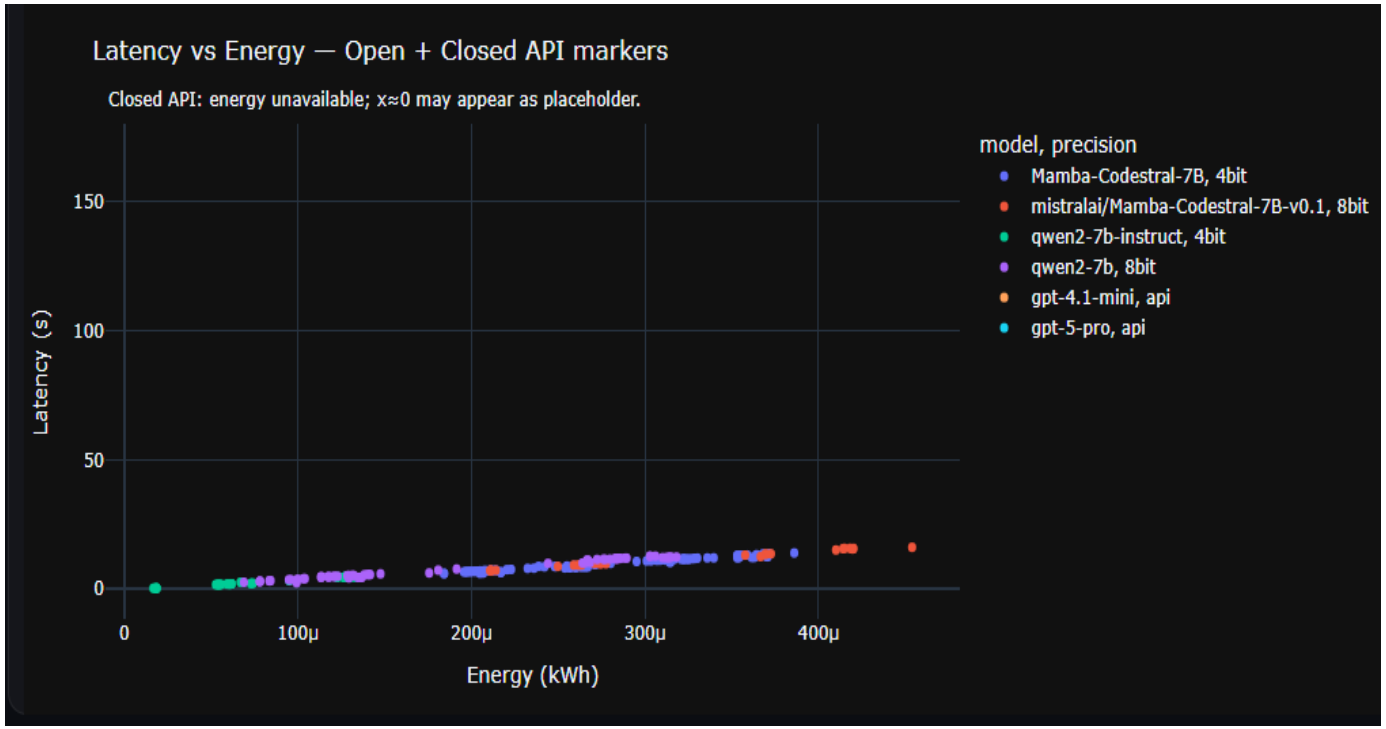


- Models were tested on **AWS GPUs (L4 & A10G)**.
- **CodeCarbon** and **NVIDIA NVML** tracked energy and CO_2.
- Tasks ranged from **Text-Generation** all the way up to **Programming/Advanced Reasoning** and **Summarization**.
- Each model ran with **4-bit and 8-bit precision**.
- All results were logged and visualized in a **Plotly Dash dashboard**.

MODELS EVALUATED

Model	Type	Precision	Key Point
Mistral-Mamba 7B	Open	4/8-Bit	Stable Baseline/ Memory-Efficient
Qwen2 7B	Open	4/8-Bit	High Throughput
GPT-4.1 Mini	Closed	API	Fast but Opaque
GPT-5 Pro	Closed	API	Accurate but Costly

ENERGY AND LATENCY FINDINGS



4-bit quantization cut energy use ≈ 35–40%.

Latency improved slightly, accuracy unchanged.

GPT-4.1 mini was fastest (≈ 2.7 s); **GPT-5 Pro** slowest (≈ 33 s).

Open 7B models balanced speed and efficiency well.

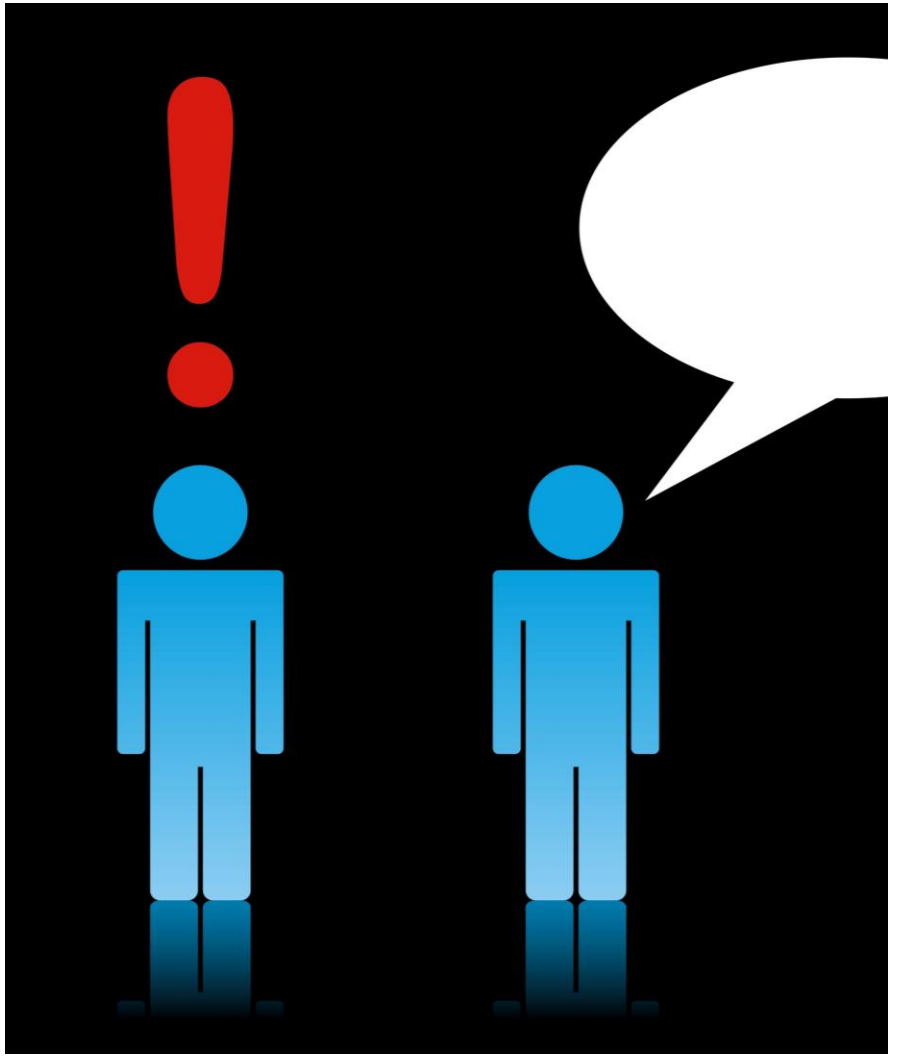
- Quantization and optimized prompts reduced power draw by up to 40% without performance loss.

INTERACTIVE LLM ENERGY DASHBOARD

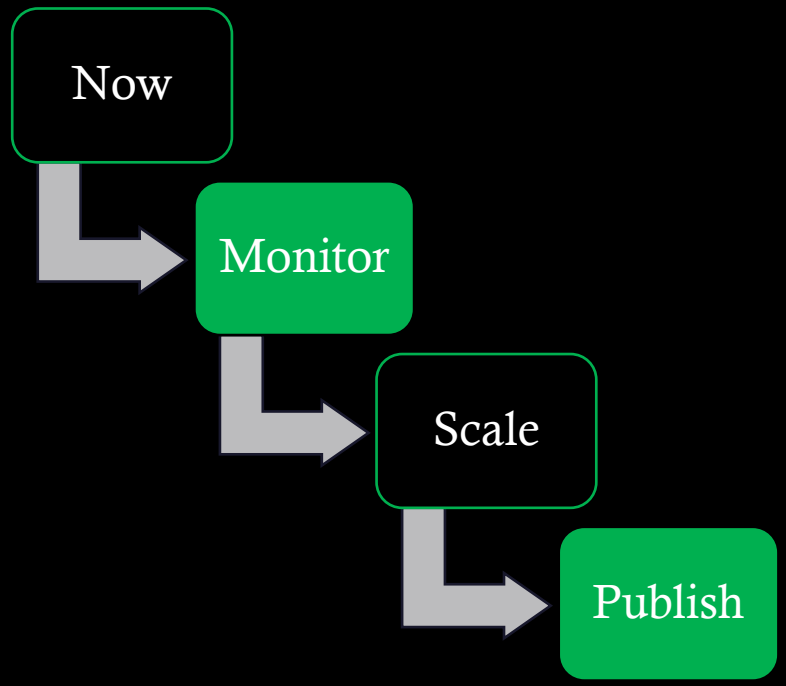
- Built in **Plotly Dash** to compare open and closed models.
- Shows energy use, latency, and CO_2 emissions side by side.
- Reveals the **transparency gap** in closed APIs.
- Preview of the Dashboard.

INSIGHTS & RESEARCH ANSWERS

Research Question	Answer
Energy Per Inference?	≈ 0.03–0.045 kWh (7B models)
Quantization/Prompt Impact?	Energy ↓ 35–40 %, Latency ↓ 10 %
Optimization Effective?	Yes, no accuracy loss
Open vs Closed?	Closed fast but opaque; Open transparent



FUTURE WORK



Implement **real-time energy monitoring agents** for inference.

Expand testing to **Jetson Orin Nano Super Developer Kit** for edge deployment.

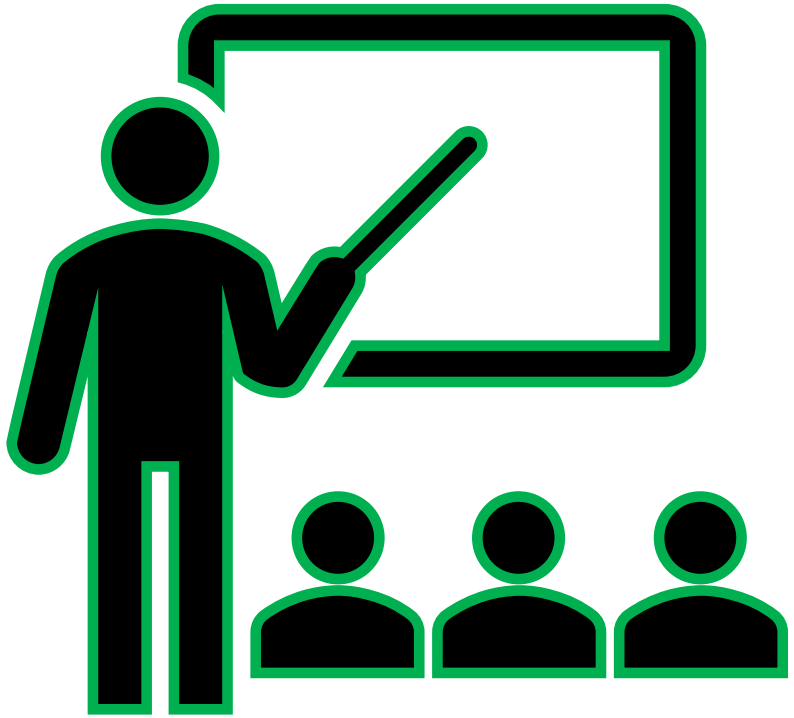
Evaluate **energy efficiency in real-world, low-power environments.**

Study GPU **temperature–efficiency relationships** under continuous load.

Develop a **public Green AI dashboard** for open benchmarking.

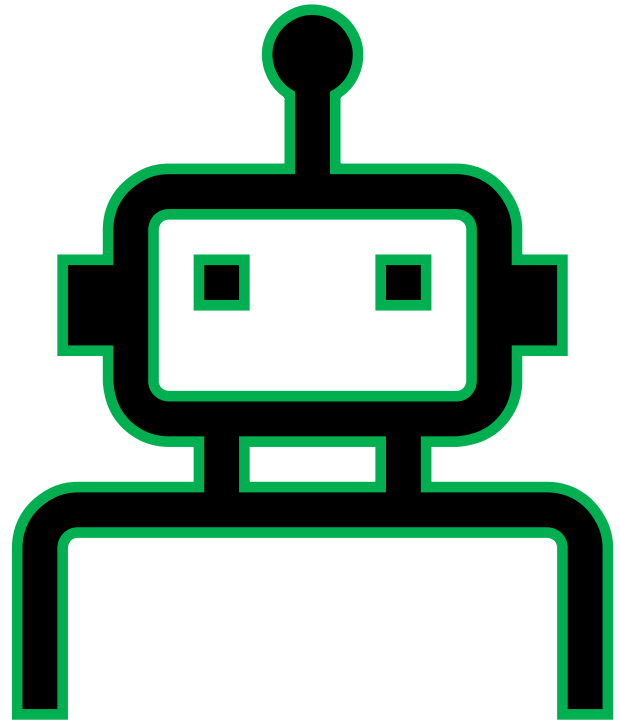
ACKNOWLEDGMENTS & Q&A

- **Instructor:** Prof. Christy Pearson
- **Mentor:** Dr. Kellen Sorauf
- **Q/A**



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



THANK YOU

Uttam Emmaniuel Dammu

- udammu@regis.edu
- https://github.com/Mrsnellek/LLM_Research_Group/tree/main/Emmaniuel