ENERGY-AWARE LLM INFERENCE ON CONSUMER HARDWARE

CSCE 585

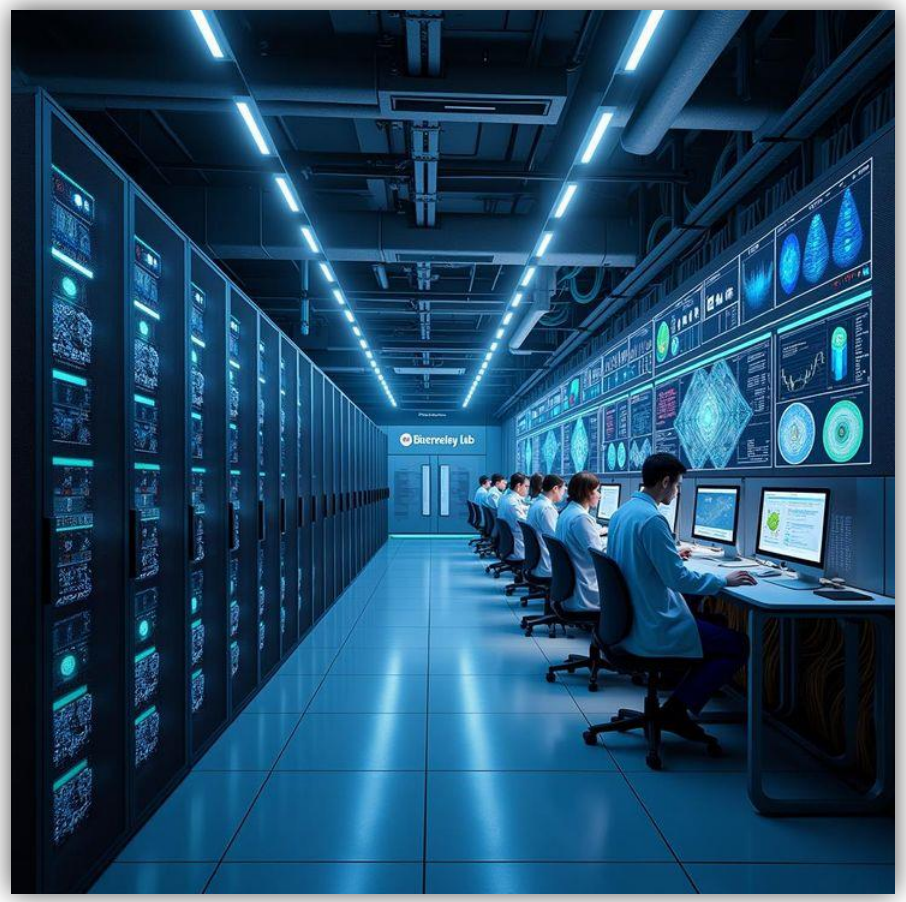
Suprawee Pongpeeradech

UNIVERSITY OF
South Carolina

Molinaroli College of
Engineering and Computing

THE PROBLEM & MOTIVATION

$650,000/year ⚡ 500,000 W



$350/year ⚡ 300 W

Molinaroli College of Engineering and Computing
UNIVERSITY OF SOUTH CAROLINA

THE METHOD

NVML (GPU)

Intel Power Gadget (CPU)

GameMaker

Intel i5 13400F + RTX 3060

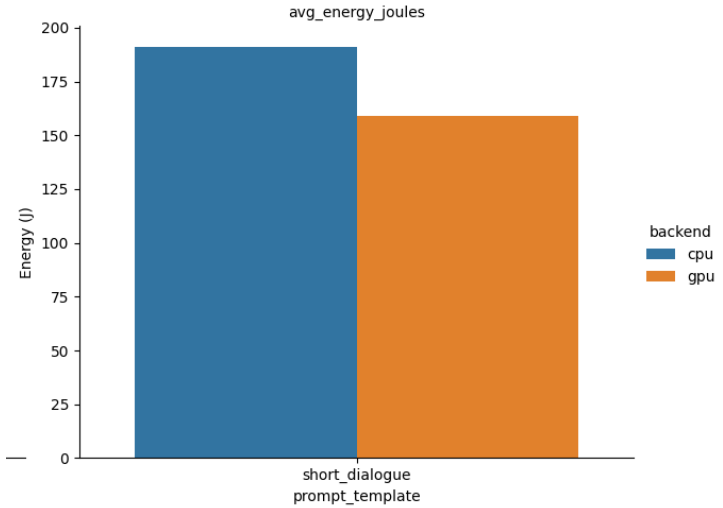
TinyLlama-1.1B

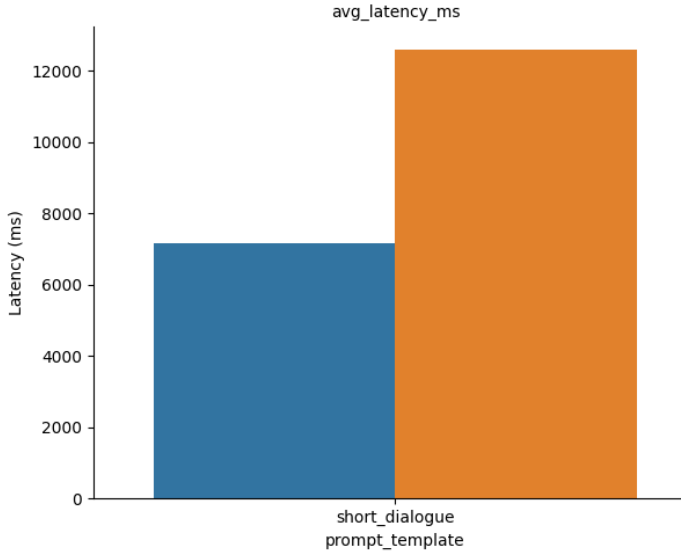
- Speed (Latency): How fast does it answer?
- Energy (Joules): How much battery would it drain?
- Efficiency (EDP): The balance between Speed and Energy.

CPU Threads: 1 vs 4 vs 8
GPU Layers: 0 (None) vs 11 (Half) vs 22 (All)
Batch Size: 128 vs 512 vs 1024

Molinaroli College of Engineering and Computing
UNIVERSITY OF SOUTH CAROLINA

EXPERIMENTS & RESULTS





Ranking (Lowest EDP is Best)

ID	Backend	Lat(ms)	Eng(J)	EDP
gpu-b1024	gpu	10925	0.00	0.00*
power_only_15	gpu	0	148.95	0.00*
gpu-l11	gpu	6562	134.31	881.37
gpu-b128	gpu	6054	154.20	933.48
gpu-l11	gpu	6415	146.91	942.37
gpu-l22	gpu	6580	156.82	1031.85
gpu-l0	gpu	6613	157.01	1038.24
gpu-l11	gpu	6329	172.39	1091.10
cpu-t4	cpu	6465	177.60	1148.29
gpu-l0	gpu	6472	179.30	1160.47

Molinaroli College of Engineering and Computing
UNIVERSITY OF SOUTH CAROLINA

DISCUSSION & INSIGHTS

- ✅ Half-GPU is the Efficiency King: Offloading 11 layers beat offloading everything. It hits the sweet spot.
- ❌ Faster ≠ Always Better: Full GPU is faster, but burns power insanely.
- 🛑 CPU Hits a Wall: Adding more than 4 threads didn't help (memory limits).
- 📉 Batch Size Doesn't Matter: For single users, it makes almost no difference.

Limitations

- I only tested short prompts (chatbots).
- Long-running heat issues weren't tested.

Molinaroli College of Engineering and Computing
UNIVERSITY OF SOUTH CAROLINA

DEMONSTRATION

https://www.youtube.com/watch?v=Q8KjrCSerJU&feature=youtu.be