GPU vs CPU Performance Analysis on Character-Level AI Models

Goda Youssif '26, Lee John '27, Sanchez Maximo '26,
Dickinson College, Department of Mathematics and Computer Science

Introduction

Modern ML workloads span a wide range of hardware varying from consumer **CPUs** to discrete **GPUs**. As small-scale model training becomes more accessible, practitioners face the question of whether available hardware is adequate or whether specific workload profiles demand specific silicon.

This study uses the **nanoGPT** framework trained on the Tiny Shakespeare **character-level** dataset as a controlled benchmark, systematically varying model complexity and context window length across five hardware configurations.

Methodology

Hardware Platforms:

Hardware	Type
NVIDIA RTX 4060	Discrete GPU
NVIDIA RTX 3070	Discrete GPU
Apple M2 Silicon	Unified Memory SoC
AMD Ryzen 7 8845 HS	x86 CPU (Modern)
AMD Ryzen 7 5800H	x86 CPU (Older)

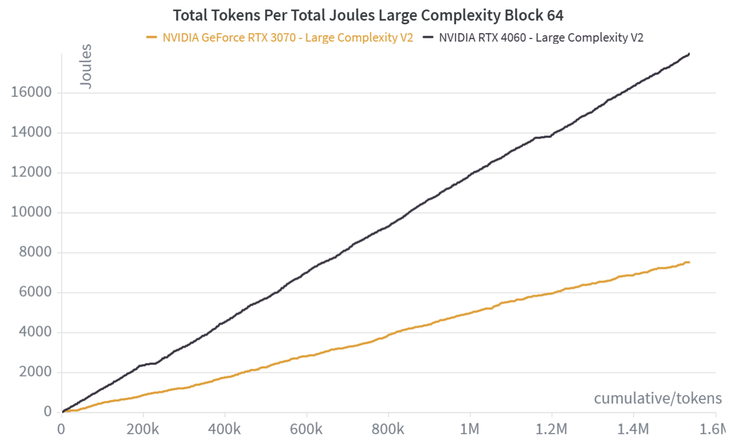
Model Complexity Levels:

Level	Heads	Layers	Embed Dim
Low	3	3	126
Mid	6	6	384
Large	12	12	768

Block Size Levels:

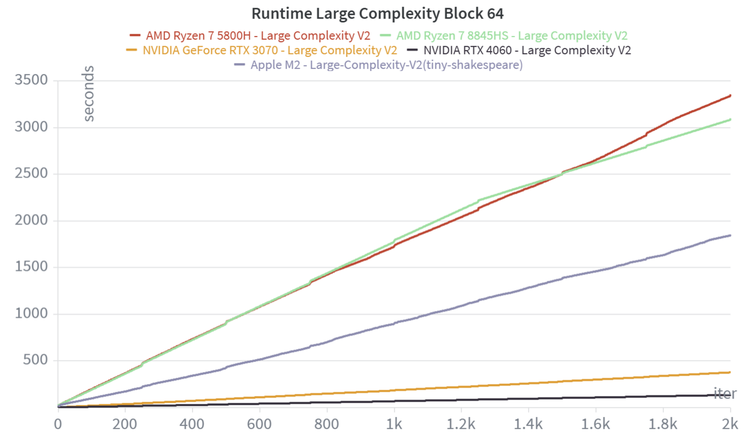
Block Size (Context Window)	Runs	Complexity Level
32	15	Low, Mid, Large
64	15	Low, Mid, Large
128	15	Low, Mid, Large

Results



RTX 4060 vs. RTX 3070 — Throughput (Tokens/sec):
★ = *winner for that configuration*

Config	RTX 4060	RTX 3070	Apple M2
Low / B32	20–30k	10–20k	25–35k ★
Mid / B32	18–24k ★	~15k	4–6k
Large / B64	12–13k ★	4–5k	~1k
Large / B128	18–25k	20–25k ★	~5k



Loss improved consistently with larger block sizes:

Block / Complexity	Low	Mid	Large
Block 32	1.977	1.928	1.992
Block 64	1.912	1.763	1.839
Block 128	1.863	1.625	1.743

Key Takeaways

- **Main finding:** GPU performance gains are workload-dependent: without sufficient model complexity or context length, the GPU is underutilized and offers limited advantage over CPUs.
- **RTX 4060:** Preferred for time-constrained training at medium-to-large complexity. Dominant in **runtime** and **throughput**.
- **RTX 3070:** Preferred under energy/thermal constraints at mid-to-large complexity. **239%** more energy-efficient at Large/Block-64.
- **Apple M2:** Competitive at low complexity due to unified memory; degrades severely at scale.
- **AMD CPUs:** Viable only at lowest workload tier. Hours vs. seconds for large-complexity tasks.
- **Block size:** Increasing block size consistently improves model quality within a fixed iteration budget.

Limitations & Future Work

- Results are based on **NVIDIA GPU-only** energy metrics, limiting cross-platform comparison
- Limited training duration (**2,000 iterations**) may not capture long-term performance effects
- Future work includes testing on **larger datasets** (e.g., OpenWebText) for more realistic benchmarks
- Explore larger model scales to analyze VRAM and **performance bottlenecks**

Acknowledgment

We would like to sincerely thank Professor John MacCormick, the Computer Science Department, and the COMP 560 student participants for their guidance, support, and collaboration throughout this project. Their insights, feedback, and contributions were invaluable in shaping our work and enhancing the overall quality of this research.

References

Gyawali, D. (2023). Comparative analysis of CPU and GPU profiling for deep learning models. arXiv:2309.02521
Li et al. (2024). Large language model inference acceleration: A comprehensive hardware perspective. arXiv:2410.04466
Li, C., Zhang, M., & He, Y. (2022). The stability-efficiency dilemma: Investigating sequence length warmup for training GPT models. arXiv:2108.06084
Li, S. et al. (2023). Sequence parallelism: Long sequence training from system perspective. ACL 2023, pp. 2391–2404