Thesis Project Portfolio

Architectural Characteristics of Apple Silicon for Machine Learning Applications

(Technical Report)

The Rise of Generative Artificial Intelligence as A Technological System

(STS Research Paper)

An Undergraduate Thesis

Presented to the Faculty of the School of Engineering and Applied Science
University of Virginia • Charlottesville, Virginia

In Fulfillment of the Requirements for the Degree
Bachelor of Science, School of Engineering

Tao Groves

Spring, 2025

Department of Computer Science

Table of Contents

Executive Summary

As artificial intelligence systems are increasingly scaled and integrated into society, their sustainability—both in terms of technical infrastructure and societal impact—is being brought into focus. This portfolio reflects a desire to understand and improve the foundations of AI, which is approached from both sociotechnical and system-level perspectives. As an exploration of the subject and a tool to target technical research, the widespread adoption of generative AI technologies is examined as a product of societal trends and policy decisions through which the evolution of AI is being shaped. Along with this general investigation, a more technical examination of performance and feasibility of computing hardware for large language models is analyzed through a comparison of architectures such as Apple Silicon and NVIDIA CUDA. This research aids in predicting the trajectory of GenAI as a technological system and supports development of efficient subsystems to improve the environmental sustainability of AI. Through these projects, a commitment is shown to ensuring that the future of AI is developed in a way that is powerful, efficient, and mindful of human and ecological needs.

Generative Artificial Intelligence (GenAI) has undergone a profound transformation in recent years from a specialized tool for optimizing infrastructure, products, and research to a ubiquitous technology shaping public discourse and consumer applications. This shift gained widespread attention in 2022 with the release of OpenAI’s ChatGPT, yet the rapid expansion of the GenAI industry was preceded by years of technological and infrastructural advancements. Since then, Large Language Models (LLMs) have become a precedent integrated into a wide range of applications, driving substantial investment and competition among companies seeking to enhance AI capabilities. However, this rapid growth has raised concerns about sustainability, as AI workloads increasingly strain global data centers and contribute to rising energy consumption. This paper explores the sociotechnical factors driving the expansion of GenAI through the lens of technological momentum, assessing how public perception and industry dynamics have shaped its development. A meta-review of Google search trends, social media engagement, financial reports, and energy consumption data is conducted to trace the interplay between societal needs and technological inertia. Additionally, a comparative analysis between the United States and China evaluates whether Generative AI is evolving into a techno-political artifact, influenced by national policies and economic strategies. By contextualizing GenAI's growth within these frameworks, insights into its past, present, and future trajectory are obtained, and practical recommendations for engineers to implement more sustainable AI systems while balancing innovation with environmental and economic concerns are offered.

Since its release in 2020, Apple Silicon's performance and role in machine learning training have attracted much attention. Compared with the previously popular machine learning training device NVIDIA GPU, Apple Silicon has significant differences in the hierarchy of storage architecture. Unlike existing training hardware frameworks that separate CPU memory and GPU VRAM, Apple Silicon uses unified memory for both the CPU and GPU. However, it is hard to tell whether unified memory is a blessing or a curse. In this paper, we select several LLM workloads and conduct end-to-end training under different memory scenarios. We conclude that the performance gap between NVIDIA GPUs and Apple Silicon is significant. We explain this performance gap on the system level by observing factors, including page fault, power consumption, kernel launch time, and thermal management. Moreover, we dive into basic linear algebra subprograms (BLAS) performance differences between NVIDIA GPUs and Apple Silicon chips to further explain this performance gap.

By exploring GenAI through both broad and specific lenses, I discovered firsthand how high-level trends in energy consumption and policy are intimately tied to low-level design decisions, such as memory architecture and kernel launch behavior. The process sharpened my ability to translate abstract sustainability goals into concrete engineering considerations and taught me to navigate the feedback loop between societal expectations and technical constraints. By integrating these perspectives, I not only gained a more holistic understanding of AI systems, but also developed a skillset that bridges technical performance analysis with policy-aware design—an interdisciplinary approach that will be crucial as AI continues to scale and embed itself more deeply into our world.