Manxi (Maggie) Shi

manxishi@mit.edu | 408-477-4595 | LinkedIn

EDUCATION

Massachusetts Institute of Technology, *BS in EECS and Physics* 4.85/5.0 **(May 2027)**

Relevant Coursework: Deep Learning (G), Hardware Arch for Deep Learning, Computer Architecture, Digital Systems (FPGAs), Accelerated Computing (GPUs) (G), Algorithms, Quantum Physics I and II, Statistical Physics, Probability, Statistical Data Analysis
Membership: IAIFI (NSF AI Institute for Artificial Intelligence and Fundamental Interactions), Society of Physics Students, SuperUROP
Skills: Python, C/C++, CUDA, SystemVerilog, RISC-V, BlueSpec, CocoTB, PyTorch

RESEARCH

SuperUROP Scholar, ***MIT GenAI Impact Consortium*** **(Jun 2025 – Present)**
Bridge statistical physics and generative models: develop a method for adaptive temperature scaling for better quality long generations in LLMs. Selected as one of the ten total MGAIC research scholars.
MIT Department of Physics, ***Prof. Soljacic*** **(Sep 2023 – Jan 2025)**
Researched novel method of light confinement in PhCs. Presented at CLEO 2025, publication in progress.

EXPERIENCE

Architecture Intern: ***Lightmatter*** **(May 2025 – Aug 2025)**
Modeled the chip's architecture, design stabilization control algorithms for on-chip photonic components.
ML Accelerator Modeling Engineer: ***Quadric*** **(Jan 2025 – Feb 2025)**
Simulated vision models running on Quadric's NPU with custom extensions to Onnxruntime library.
Software Engineer Intern: ***Amlogic Inc.*** **(June 2024 – Aug 2024)**
Performance engineered audio resampling algorithms with Neon intrinsics on ARM architecture.

HONORS

2022 US Physics Team Member (Top 20 in the US), US Physics Olympiad Gold Medalist (2022)
3x AIME Qualifier (2021/2022/2023)

PROJECTS

FPGA Implementation of Optical Flow (6.205 Digital Systems) (Oct 2025 – Dec 2025)
Designed and implemented motion tracking with optical flow on an FPGA with real-time camera input.
GPU Kernels (6.s894 Accelerated Computing) (Sep 2025 – Dec 2025)
Hand-optimized CUDA kernels for various workloads, learned about architecture of GPUs and TPUs.
Investigating the Impact of Accelerators and Extreme Sparsity (March 2025 – May 2025)
Designed and benchmarked accelerator architectures to optimize sparse matrix computations.
A Study of Variation Across Attention Heads and Layers (Nov 2024 – Dec 2024)
Proposed novel method of weight sharing for efficient attention computation. Blog link.
MIT Lincoln Laboratory (June 2020 – July 2020)
Private automated contact tracing with Raspberry Pi's and machine learning for COVID-19: piPACT.