Malav Patel

mpatel636@gatech.edu | malav-p.github.io | github.com/Malav-P

Skills

Languages: Python, C, C++, TypeScript, CUDA, Julia, MATLAB, SQL
Frameworks: PyTorch, Tensorflow, JAX, sklearn, OpenCV, Docker, Kubernetes, AWS, FastAPI, PostgreSQL, Slurm, Node, CI/CD
Research Topics: Diffusion, NLP, Filtering, Constrained Optimization, GNC, HPC

Experience

Research Engineer, Georgia Institute of Technology – Atlanta, GA August 2022 – Present
- Applying gradient-based optimization, integer programming, and filtering to optimize placement and tasking of satellite constellations in cislunar space.

Founder, ml-code – Atlanta, GA January 2026 – Present
- Building a full-stack containerized service for AI-assisted machine learning pipeline generation, execution, and monitoring. FastAPI backend on AWS EC2 + CLI + TypeScript frontend.

AI Trainer, Data Annotation – Atlanta, GA June 2023 – January 2025
- Engineering difficult prompts and problems to challenge SOTA multi-modal models across english, mathematics, and science tasks. Effected 1.3% improvement in average model accuracy.

Machine Learning Engineer, Startup – Atlanta, GA August 2025 – December 2025
- Finetuning Vicuna7B for text/image-conditioned floorplan generation tasks. Reduced memory footprint by 70% via LoRA and QLoRA techniques, enabling training on consumer grade GPUs.
- Direct Preference Optimization to align model on layout preferences, yielding 34% improvement in quality.

Projects

On Device GPT Inference github.com/Malav-P/staticgrad
- Built a library in C++ for running training and inference of GPT2.
- Optimized inference time on Apple M1 Hardware using Apple's Accelerate GEMM kernels and custom attention kernel, achieving inference time latency of 25 ms/token without compiler optimizations.

Open SoundStream malav-p.github.io/soundstream
- Established an open source implementation of SoundStream, a neural audio codec underpinning SOTA generative audio models in 2 weeks.
- Trained on a single HPC NVIDIA L40S GPU node in less than 15 hours. Released checkpoints for other users.

ModernBERT for Patent Classification github.com/Malav-P/modernpatentBERT
- Trained modernBERT on the patent classification task with >2x speedup in training throughput achieving SOTA precision and F1 scores on a domain test set.
- Hosted an open repository on Huggingface of 3 million patents for future work on the classification task.

CNNs in C++ github.com/Malav-P/CNN
- Built a C++ header-only library for constructing, training, and running convolutional neural networks.
- Achieved hand-written digit recognition at 60 Hz on Apple M1 with a CNN trained on MNIST.
- Implemented custom matmul kernels in CUDA to parallelize intensive matrix multiplication operations, achieving a 3x speedup on matmul calls.

Education

Georgia Institute of Technology – PhD in Aerospace Engineering June 2026
Georgia Institute of Technology – MS in Computer Science December 2025
Northwestern University – BS in Physics, Mechanical Engineering, Integrated Sciences June 2022