

Phong Cao

(+1) 774-701-3932 phongct1105@gmail.com linkedin.com/in/phong-cao github.com/PhongCT1105

EDUCATION

Worcester Polytechnic Institute

Expected May 2027

BS in Computer Science; MS in Artificial Intelligence

Cumulative GPA: 3.92/4.0

- **Coursework:** Algorithms, Operating Systems, Databases, Machine Learning, Deep Learning, NLP, AI Accelerators
- **Honors:** Dean's List all semesters, Presidential Scholarship, National Science Foundation REU
- **Awards:** \$10K+ prize value; Runpod Grand Prize, Stanford AI Hackathon 3rd Place, UC Berkeley AI Hackathon 2nd Place, NASA Space Apps People's Choice, WPI GoatHacks 2x Winner

SKILLS

Languages: Python, TypeScript, JavaScript, Java, C, C++, SQL, Bash, C#, Kotlin

Backend & Frontend: FastAPI, Flask, Node.js, Express.js, React, Next.js, Prisma, REST APIs, SSE, Auth0, OAuth 2.0, JWT

Databases & Cloud: PostgreSQL, Supabase, MongoDB, Redis, Pinecone, ChromaDB, AWS, Azure, GCP

DevOps & Observability: Docker, Kubernetes, Terraform, GitHub Actions, pytest, Sentry, OpenTelemetry, Git, Linux

AI/ML Systems: PyTorch, scikit-learn, Hugging Face, ONNX Runtime, TensorFlow, CUDA, LangChain, OpenAI SDK

EXPERIENCE

Founder & Software Engineer

May 2026 – Present

Zolli Labs – FlashML Distributed ML Platform

San Francisco, CA

- Founded **FlashML**, an open-source, self-hostable **distributed ML platform** that orchestrates jobs across heterogeneous compute and resumes interrupted work after worker failures; won the Runpod Grand Prize and \$8K.
- Enabled safe compute sharing through a CLI-based **Docker host agent** that sandboxed ML jobs; blocked **24 adversarial file, network, and resource-exhaustion attacks** while peaking below **0.25% of memory** on a 16 GB host.
- Cut batch completion time by **47%** versus static scheduling on the same workers by engineering a **pull-based distributed scheduler** that let faster machines claim more jobs across a **3.7×** speed range.
- Shipped FlashML to 100+ pilot users as **Python packages on PyPI**, automating tests, builds, and releases via **GitHub Actions CI/CD**; delivered **PyTorch, scikit-learn, Hugging Face** adapters backed by **2,200+ tests** in pytest.

Research Assistant – AI/ML Infrastructure

Jun 2025 – Dec 2025

National Science Foundation

Worcester, MA

- Co-authored and implemented **LLM-TSFS**, an LLM-guided selector over 157 exogenous features for data-scarce time-series forecasting, reducing RMSE by up to **27%** versus an XGBoost baseline; presented at **IEEE MIT URTC** conference.
- Cut experiment setup from hours to minutes by engineering a config-driven **Python/Bash/YAML HPC launcher** that automated job submission, result collection, and failure triage across 400+ reproducible runs.
- Grounded LLM feature selection in current domain evidence using a **Hugging Face/ChromaDB RAG pipeline** with cross-encoder reranking, cutting forecast RMSE by up to **60%** versus no RAG.

Software Engineer Intern – Applied AI

Jun 2024 – Aug 2024

FPT Software

Vietnam

- Built an **Azure-based multimodal retrieval platform** indexing 5,000+ event photos, enabling non-technical users to find semantically matched images through natural-language queries instead of manually scanning albums.
- Achieved **97% Recall@5** accuracy with 300ms query latency by fine-tuning and benchmarking an **Azure AI Vision** retrieval pipeline backed by Blob Storage and Cosmos DB for scalable image and metadata indexing.
- Converted PoC approval into a **\$30K follow-on contract** by shipping web and Android apps with **Python, C#, Kotlin**.

AI Product Tester & Campus Ambassador

Feb 2026 – Present

Adobe

Remote

- Partnered with product managers and technical leads to test **Adobe Acrobat Student Spaces**, tracing high response latency to redundant AI tool calls and submitting reproducible failure cases for product refinement.
- Invited back as a Campus Ambassador after the testing program, driving product adoption among **100 WPI students**.

Teaching Assistant – Software Engineering

Mar 2024 – May 2024

Worcester Polytechnic Institute – Mass General Brigham Collaboration

Worcester, MA

- Led 11 students across 5 Agile sprints to deliver a full-stack hospital navigation and service platform, guiding API design, database architecture, sprint planning, code reviews, and stakeholder demonstrations.
- Built **Express/Prisma** routing APIs over 147 kiosks and deployed the PERN application on **AWS** with PostgreSQL, Auth0/JWT access control, and Dijkstra, A*, BFS, and DFS pathfinding.

PROJECTS

On-Device Real Estate Q&A Agent | Edge AI Deployment | ONNX Runtime, Android ARM64 | GitHub

May 2026

- Built a real-estate Q&A assistant for offline **edge inference** on Android ARM64 by fine-tuning **Hugging Face FLAN-T5**.
- Exported the model to **ONNX** and built Android inference with **ONNX Runtime** and native **C++ SentencePiece**.
- Optimized FLAN-T5 for ARM64 deployment across 14 **FP16/BF16/INT8 quantization** and structured/unstructured pruning variants; selected FP16 to cut size **50.5%** and latency from **5.48s to 163ms** while retaining **96.6% Token F1**.