My Le

(217) 419-9649 | iamlhhm@gmail.com | linkedin.com/in/lehoanghamy | github.com/le-hoang-ha-my

Education

Case Western Reserve University (Combined Bachelor's/Master's Program) Jan. 2026
Master of Science in Computer Science (Artificial Intelligence) ***GPA:*** *4.0/4.0*

Case Western Reserve University Jan. 2026
Bachelor of Science in Computer Science (Artificial Intelligence), Computer Engineering ***GPA:*** *3.7/4.0*

- **Minors:** Mathematics, Electrical Engineering

Experience

Machine Learning Research Assistant Mar. 2024 – Present
Ray's AI Lab, Case Western Reserve University *Cleveland, OH*

- Developed a sparse dictionary learning algorithm for neural touch response reconstruction in PyTorch, achieving >95% classification accuracy with 80% dimensionality reduction.
- Optimized regularization strategies for tactile stimulus reconstruction in prosthetic applications, maintaining >80% F1 score on reconstructed neural data.

Teaching Assistant – CSDS 440: Machine Learning Aug. 2024 – Dec. 2024
Dept. of Computer and Data Sciences, Case Western Reserve University *Cleveland, OH*

- Evaluated assignments and provided technical feedback on graduate-level coursework for 30+ students.
- Conducted support sessions to debug PyTorch implementations and clarify concepts.

Research Assistant Aug. 2022 – Mar. 2024
ERIE Lab, Case Western Reserve University *Cleveland, OH*

- Co-developed an algorithm for telesurgical force estimation using crowd-sourced labels and robot sensor data to bypass ground truth force sensors in clinical settings.
- Trained custom CNNs and fine-tuned EfficientNetB3 in PyTorch on NVIDIA V100 GPUs via SLURM, achieving >90% contact detection accuracy and <10% force prediction error across 160K+ surgical video frames.

Projects

AI-Powered Financial Document Processing Platform | *Next.js, Supabase, Vercel AI SDK, Tailwind CSS*

- Built and deployed a multimodal RAG system with Next.js 15, Tailwind CSS, and Gemini 1.5 Flash, achieving 95%+ field extraction accuracy and <3s latency on 1-5 page financial documents via parallel API calls.
- Implemented an agentic query router using Vercel AI SDK, improving retrieval relevance by 40%.
- Secured multi-tenant architecture with Supabase Row Level Security and optimized React Server Actions with connection pooling, achieving <200ms p95 query latency.

Autonomous GPU Cluster Orchestrator | *Python, Docker, AWS, LangGraph, Streamlit, Prometheus*

- Built a scheduling agent with LangGraph and Qwen2.5-7B-Instruct that increased GPU utilization by 20% via dynamic bin-packing and priority-aware preemption, reducing P0 job wait times by 45% over FIFO queues.
- Deployed distributed monitoring with Prometheus across AWS ECS to track containerized workloads, automatically detecting and terminating zombie Docker processes to reclaim 3GB+ VRAM per node daily.
- Developed a real-time Streamlit dashboard displaying live GPU telemetry, node health metrics, and agent decision traces with sub-second refresh rates.

Publications

Learning Low-dimensional Local Features from Somatosensory Neural Data
IEEE Transactions on Neural Systems and Rehabilitation Engineering, 2026 (first author) (In Review)

Vision-Based Force Estimation for Minimally Invasive Telesurgery Through Contact Detection and Local Stiffness Models
Journal of Medical Robotics Research, 2024 (second author), IROS 2023 Poster (first author)

Technical Skills

Languages: Python, Java, TypeScript, JavaScript, HTML, CSS, SQL
Tools & Frameworks: PyTorch, Git, GitHub, Linux/Unix, Next.js, React, Node.js, PostgreSQL, Scikit-learn, Supabase, AWS, Docker, LangGraph, LangChain, Vercel, HPC, Prometheus