Mohammad Aaron Chegini

+1 (250) 415 6325 | mchegini@uvic.ca | github.com/moh-c | linkedin.com/in/moh-chegini/

Machine Learning Engineer delivering end-to-end AI solutions. My expertise is in turning complex requirements into production-ready LLM and Computer Vision systems, demonstrated by architecting a full-stack AI paralegal and deploying an 8x H100 GPU cluster. I thrive on using Python, PyTorch, and AWS to create high-impact products.

SKILLS

Machine Learning & AI	LLMs, RAG, Fine-Tuning (LoRA, QLoRA), Prompt Engineering, Computer Vision (YOLOv8), PyTorch, TensorFlow, Hugging Face, Reinforcement Learning
Infrastructure & Cloud	AWS (EC2, RDS, S3), Terraform, Docker and Docker Compose, Git, MLflow HPC Cluster Configuration (NVIDIA H100)
Languages & Libraries	Python, C/C++, JavaScript, MATLAB, NumPy, Pandas
Full-Stack & Edge	Django, React, Jetson Orin Nano, Raspberry Pi, ESP8266, Arduino

EXPERIENCE

ServiceNow, ***ML Research Scientist*** *(Collaboration)* | Vancouver, BC — Jul 2024 - Jan 2025

- Co-authored InsightBench, a novel benchmark for evaluating business analytics agents, leading to a publication at ICLR 2025.
- Developed and benchmarked multi-step LLM agents for complex business analytics, contributing to ServiceNow's AI research initiatives.

Royal BC Museum, ***LLM*** *Engineer* | Victoria, BC, Canada — Nov 2024 - Jan 2025

- Enabled real-time, interactive AI exhibits by integrating OpenAI's API via WebSockets and a custom Unity-compatible backend.
- Improved museum-specific AI response accuracy and reliability by implementing GraphRAG, ensuring a seamless visitor experience.

Innovation Support Unit, ***LLM*** *Engineer* | UBC, BC, Canada — Nov 2023 - Jan 2024

- Reduced model hallucinations in medically-sensitive exercise suggestions by 40% by implementing a RAG system with Llama-2.
- Designed and optimized a context-aware recommendation system (weather/time/safety) using GPT-4, Mistral, and Mixtral for reliable activity outputs.

Prism Technology Holdings, *Python Developer* | Remote (USA) — Jan 2022 - Apr 2022

- Cut real-time analysis processing time by optimizing Python DSP algorithms with multi-threading.
- Enhanced chest movement detection accuracy for a mindful breathing app by developing custom Digital Signal Processing (DSP) algorithms.

NikTech Org., *Lead Fullstack Engineer* | Tehran, Iran — Jul 2018 - Dec 2019

- Led a national project for Iran's Roads Ministry, deploying a CNN-based erosion detection system at a major international airport.
- Engineered a full-stack Django/React app with OpenLayers for real-time erosion mapping, streamlining maintenance workflows.

PROJECTS

AI Legal Assistant (Persian Language) | Jan '25 - Present

- Architected and built a full-stack AI paralegal, fine-tuning an LLM on a custom-crawled corpus of Persian legal documents.
- Engineered a scalable data pipeline using Python and managed infrastructure as code with Terraform on AWS (EC2, RDS).
- Implemented a cached OpenAI API layer, significantly reducing operational costs.

Real-time Movement Analyzer | Jan '25 - Present

- Developing a high-performance computer vision system for real-time movement analysis using YOLOv8 and MediaPipe for object detection and advanced pose estimation.
- Optimizing models for deployment on edge hardware (Raspberry Pi), achieving near-real-time processing for interactive feedback.

HPC Cluster Deployment (for NC A&T University Lab)

- Solely configured an 8x NVIDIA H100 GPU compute cluster, enabling the lab to conduct large-scale AI research and model training.
- Managed and helped significant aspects of the setup, from hardware installation and network configuration to the software environment.

AaronNet for ITU ML/AI challenge in 5G

- Developed AaronNet, a novel, quantized neural network for RF modulation, achieving 64x lower inference cost for edge deployment.
- Achieved state-of-the-art accuracy (surpassing VGG baseline), securing 3rd place in the global ITU ML/AI in 5G Challenge.

ML-Accelerated Robotic Arm Kinematics

- Developed a Neural Network to approximate inverse kinematics for a 3-DoF robotic arm, enabling faster motion planning than traditional solvers.
- Utilized Docker Compose for containerized deployment and architected the system for future extension to 6-DoF.

EDUCATION, PUBLICATIONS & HONORS

University of Victoria, Victoria, BC, Canada — May 2023 – Present

- MASc in Electrical & Computer Engineering, CGPA: 4.0/4.0

Shahid Beheshti University, Tehran, Iran — Sep 2018 – Mar 2023

- Bachelor's in Electrical Engineering, GPA: 17.9/20.0

ITU ML/AI in 5G Challenge | Global Competition — December 2021

- Awarded 3rd Place Globally among 1600+ teams from industry and academia.

Key Publications | Top-Tier Venues — 2020 – Present

- Authored/co-authored 7 papers in venues including ICLR 2025 and IEEE.
- Full publication list available on Google Scholar.