Alireza Mounesisohi, Ph.D.

AI Agentic Systems Expert | Data Scientist & Machine Learning Engineer

Los Angeles, CA | (925) 239-9892 | US Citizen | armsohi@gmail.com | linkedin.com/in/mounesi | github.com/mounesi

Expert Profile

- AI/ML expert with a Ph.D. in AI and Robotics and 8+ years building and operating production artificial-intelligence systems across healthcare, biotech, enterprise, and aerospace. Specialized authority in **agentic AI** — the design, deployment, and governance of autonomous, tool-using LLM agents — with hands-on production experience across the leading agent frameworks (LangChain, LangGraph, CrewAI, and the Model Context Protocol / MCP).
- **Available for expert consulting, technical advisory, and expert-witness engagements** on: agentic AI architecture and multi-agent orchestration; LLMs (RAG, fine-tuning, evaluation, safety); LLM inference infrastructure and GPU deployment (vLLM, NVIDIA H200, NVAIE/NIMs); AI in regulated environments (FHIR, CMS, HIPAA-aligned deployments, AI compliance); machine learning, computer vision, and autonomous robotics; and AI ethics and responsible-AI governance.

Areas of Expertise

Agentic AI & Orchestration:	Autonomous and multi-agent systems; tool/function calling; LangChain, LangGraph, CrewAI; MCP server design; agent evaluation, guardrails, and reliability
Generative AI & LLMs:	RAG; LLM fine-tuning and model lifecycle; open-weight model deployment (Kimi K2.5, GPT-OSS, Llama Scout/Maverick); Azure OpenAI, Amazon Bedrock
AI Infrastructure & MLOps:	vLLM inference; NVIDIA H200, NVAIE, NIMs, NVIDIA AI Factory, Omniverse; Kubernetes (EKS) for HA training/inference; quantization and GPU acceleration
AI in Regulated Domains:	Secure healthcare data exchange (FHIR, CMS-aligned interoperability); HIPAA-compliant AI deployments; cross-organizational security and compliance
ML, Vision & Robotics:	Deep learning, NLP, and vision systems; autonomous robotics in unstructured environments; large-scale ML/NLP data pipelines

Professional Experience

Immunity Bio *Apr. 2025 – Present*

Data Scientist & Machine Learning Engineer (hybrid) Los Angeles, CA

- Provisioned and operate vLLM inference on NVIDIA H200 infrastructure for open-weight models including Kimi K2.5, GPT-OSS, Llama Scout, and Llama Maverick, integrated with NVIDIA AI Enterprise (NVAIE), NIMs, and NVIDIA AI Factory patterns for scalable GenAI delivery.
- Architect agentic and RAG-oriented workflows with LangChain, LangGraph, CrewAI, and MCP servers for tool-calling and orchestration; support model lifecycle and fine-tuning; align automation with NVIDIA Omniverse.

Dock Health *May 2023 – Mar. 2025*

Machine Learning Engineer (Remote) Boston, MA

- Partnered with external healthcare customers and integration stakeholders to ship AI capabilities (including patient synopsis and RAG-backed flows) into production, coordinating security, HIPAA, and operational requirements across organizations.
- Developed an AI-based patient synopsis on Amazon Bedrock on EKS; managed Kubernetes for high availability and scale.

Union.ai *Aug. 2022 – Apr. 2023*
Lead AI Solutions Engineer (Remote) Seattle, WA

- Built large-scale ML and NLP pipelines in AWS (Python, Go) for external biotech customers including Delve Bio, AbCellera, and Cradle, with NLP-driven extraction and interpretation of biological and trial data.
- Cut experiment initiation from 10+ minutes to under one minute for metagenomic sequencing and clinical-trial workflows.

Domino Data Lab *Oct. 2021 – Jun. 2022*
Field MLOps Engineer San Francisco, CA

- Supported **external enterprise customers** on distributed AI and MLOps: optimized deep learning via quantization and NVIDIA GPU acceleration; deployed and maintained Kubernetes for resilient training and inference pipelines.
- Collaborated with Bristol Myers Squibb on benchmarking distributed computing for an AI pipeline labeling chemical images (results presented at IEEE EMBC 2022).

HOME Space Technology Lab (NASA-funded) *Mar. 2020 – Sep. 2021*
Software Engineer (Machine Vision and Robotics) Davis, CA

- Developed ML-based robotic algorithms for critical space habitat tasks and integrated a deep learning vision system for mobility in unstructured environments.

AHMCT, UC Davis *May 2017 – Sep. 2021*
Software Engineer (CALTRANS project) Davis, CA

- Built an autonomous robotic system with computer vision for real-time asphalt defect detection (70× speedup) and a control interface with real-time deep learning analytics.

Education

University of California, Davis *2017 – 2022*
Ph.D., AI and Robotics Davis, CA

San José State University *2015 – 2017*
M.S., Mechanical Engineering San José, CA

University of Science and Technology *2010 – 2014*
B.S., Mechanical Engineering Tehran, Iran

Selected Projects

Insight Data Science *Jun. 2020 – Aug. 2020*
Data Science Fellow New York, NY

- Deep learning for pavement assessment using Google Street View data; potential savings up to $200,000 per US city. Code: github.com/mounesi/pa.

Center for Information Technology Research in the Interest of Society *Sep. 2019 – Feb. 2020*
Software project (CITRIS) Davis, CA

- AI/VR toolkit for VR puzzle applications; C#/Unity front-end with three puzzle apps.

Advisory & Service

Merritt College *Nov. 2023 – Present*
Advisor Committee Oakland, CA

- Advise on AI ethics in business education; foster CIS and Marketing collaboration; integrate AI into finance and business courses.

Technical Skills

Agent Frameworks:	LangChain, LangGraph, CrewAI, MCP servers, agentic and multi-agent workflows
GenAI / LLMs:	RAG, LLM fine-tuning, vLLM, Azure OpenAI, Amazon Bedrock, open-weight model deployment
NVIDIA / Infra:	NVIDIA H200, NVIDIA AI Enterprise (NVAIE), NIMs, NVIDIA AI Factory, Omniverse, CUDA ecosystem
Cloud & DevOps:	AWS (ECR, IAM, S3, Lambda, Glue, Redshift, Fargate), Azure, GCP, Kubernetes (EKS, Helm, RBAC), Docker, Terraform, CI/CD (GitHub Actions, Jenkins)
Data / ML:	FastAPI, Apache Spark, Hadoop, Airflow, PyTorch, TensorFlow, Keras, scikit-learn
Languages:	Python, SQL, C++, Bash, Java, Scala, JavaScript, R, C#, HTML/CSS
Data stores:	PostgreSQL, MongoDB, MySQL, Redis

Publications

- Jennifer D., Alireza M., & Marcin M. (2022). Benchmarking Distributed Computing for an Artificial Intelligence Pipeline for Labeling Chemical Images. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC 2022)*. IEEE.
- Alireza Mounesisohi (2021). *Sensor Directed AI-Based Robot Task Planning in Unstructured Environments.* Ph.D. dissertation, UC Davis.
- Bennett, D., Mounesisohi, A., Swanston, T., & Ravani, B. (2020). *Research to Support Crack Cleaning Operations in Moving Lane Closures* (Report No. CA19-3176).
- Alireza Mounesisohi (2017). Control System Design for Disturbance Rejection in Active Vibration Control System. M.S. thesis.
- Mounesisohi, A., & Bashash, S. (2017, August). Vibration compensation of display contents in smart devices using accelerometer feedback. In *2017 IEEE Conference on Control Technology and Applications (CCTA)* (pp. 420–425). IEEE.