Hosein Hojat Ansari

LLM & MLOps Engineer

linkedin.com/in/hhojatansari

github.com/hhojatansari

Tehran, Iran [1992-11-16]

hhojatansari@gmail.com

0098 911 648 2390

About Me

In recent years, I have focused on MLOps and AI Agents, ensuring that machine learning systems and LLMs are reliably deployed and operated in production environments. This has involved working with event-driven workflows, automated pipelines, and scalable deployments.

Currently, I am expanding my focus toward the DevOps domain. I am particularly interested in deployment, automation, and building robust infrastructure, and I am motivated by the challenge of improving how systems are delivered and maintained in real-world environments.

Education

Bachelor of Electronic Engineering

Islamic Azad University of Lahijan *Lahijan, Guilan, Iran*

Master of Artificial Intelligence

Islamic Azad University of Lahijan *Lahijan, Guilan, Iran*

Work Experience

MLOps Engineer

Ertebat Farda co

Oct 2025 - Present *(On-site) - Tehran, Iran*

AI Platform & API Development

- Designed and developed a high-performance API Gateway for the company's chatbot engine using FastAPI, serving as the main entry point for AI services.
- Implemented Server-Sent Events (SSE) to support real-time, streaming responses for conversational AI use cases.
- Focused on scalable, production-ready API design with clear separation between gateway logic and AI service backends.

Infrastructure, Kubernetes & Deployment

- Installed and configured a Kubernetes cluster to host and manage AI services in a production environment.
- Deployed multiple AI and backend services on Kubernetes, enabling improved scalability, isolation, and operational stability.
- Worked closely with containerized workloads to ensure reliable deployments and smooth service orchestration within the cluster.

MLOps Engineer

HARA AI

April 2022 - Aug 2025 (3 year 5 month) *(Remote) - Tehran, Iran*

AI Model Development & Optimization

- Researched and experimented with advanced loss functions such as **OSM** to improve **YOLOv5 object detection** accuracy, gaining valuable experience in model optimization techniques.

Infrastructure & ML/DevOps

- **Containerized** all AI services, handling complex dependencies to ensure smooth and reproducible deployments using **Docker**.
- Designed and implemented **GitLab CI/CD pipelines** to fully automate the deployment process — building images on each `main` branch commit, testing in staging, and deploying to production environments.
- Set up a **centralized monitoring and observability platform** using **OpenTelemetry** and **SigNoz**, instrumenting all services to collect logs, metrics, and traces in one place.
- Developed a **Rocket.Chat bot** to send real-time alerts whenever a service encountered issues, significantly reducing downtime and response time.
- Built a **CLI package** to automate repetitive tasks such as:
    - Generating standardized AI service templates.
    - Deploying models from **MLflow registry**.
    - Providing unified logging with support for both local files and **OpenTelemetry/SigNoz**.

System Integration & Workflow Design

- Developed a **Python-based RPC interface** to connect the AI backend with the company's **Java API**, ensuring reliable inter-service communication.
- Built a **real-time camera management service** for the **facial recognition attendance system**, which:
    - Dynamically detected newly registered IP cameras from the database.
    - Initialized and managed camera streams with **OpenCV**.
    - Streamed frames to face detection and recognition services using **gRPC**.
- Designed an **async, event-driven workflow** for the **call center QC project** using **LangGraph**, enabling horizontal scaling by increasing service instances.
- Created **interactive Dash/Plotly dashboards** for visualizing data and testing various AI services.

Machine Learning Engineer

TabiateZendeh Laboratories (Cinere)

Jul 2020 - Mar 2022 (1 year 9 month) *(on-site) - Tehran, Iran*

Predictive Modeling & Business Insights

- Developed and deployed **time series forecasting models** for product sales prediction, supporting strategic business planning and inventory management.
- Built and launched **customer churn prediction models** to forecast three-month churn risk, providing actionable reports to the sales team to improve retention strategies.

Dashboards & Reporting

- Designed and implemented a **comprehensive AI dashboard** visualizing:
  - Sales forecasts
  - Churn probabilities
  - Sales performance vs. targets
- Built an **automated reporting system** that generated and sent scheduled reports to sales managers.

Image Processing & Data Engineering

- Applied **OpenCV image processing techniques** to analyze laboratory vial images for R&D purposes.
- Built **web crawlers** using **Selenium** to collect, clean, and structure data from various websites for internal analysis.

Infrastructure & Automation

- Managed **Linux server configuration and updates**, ensuring the stability and availability of ML services and tools.
- Developed and deployed **Flask APIs** to serve AI models and integrate them with other company systems.
- Automated repetitive tasks and workflows using **custom scripts** and **cron jobs**, improving operational efficiency.

Machine Learning Engineer

GATA, TOSAN Holding

Jul 2020 - Mar 2022 (1 year 1 month) *(on-site) - Tehran, Iran*

Computer Vision & Model Development

- Developed lightweight, high-performance **object detection models** using **OpenCV Haar Cascade**, optimized for real-time applications.
- Trained and fine-tuned **license plate detection models** using **TensorFlow API**, delivering a fast, accurate, and mobile-friendly solution for deployment on embedded devices.
- Researched and experimented with multiple **text detection models** (e.g., **PixelLink**) to evaluate performance for real-world use cases.

Data Generation & Crawling

- Built **synthetic dataset generation** using **OpenCV**, producing highly realistic data for:
  - **French vehicle license plates**
  - **Iranian national ID cards**
  - Included advanced transformations such as noise, perspective warping, and rotation to closely mimic real-world conditions.
- Developed large-scale **Instagram image crawling pipelines**, automatically collecting videos containing faces based on hashtags and filtering them with computer vision techniques.

Deployment & Embedded Systems

- Ported AI software written in **C++** to run on various **Linux distributions** and architectures by rebuilding and recompiling dependencies (including **TensorFlow** from source).
- Deployed and optimized **license plate recognition models** on **embedded devices** such as **Raspberry Pi** and **NVIDIA Jetson** for edge computing scenarios.

Software Development & Testing

- Developed **RESTful APIs** using **Flask** to make AI services production-ready and easily consumable by other systems.
- Designed and implemented **unit and integration tests** for AI services to ensure reliability and stability across deployments.

[Programming Language]

Python	Proficient (daily use, primary language)
C++	Experienced (previous professional use, less recent)
Bash	Experienced (scripting, automation, deployment)
C#	past projects

[DataBase]

MongoDB

My SQL

Postgres

[Tools & Frameworks]

Version Control	git
Deployment & Messaging & IaC	Docker, GitLab CI/CD, RabbitMQ, gRPC, Ansible, Terraform
Monitoring & Observability	OpenTelemetry, SigNoz, Grafana, Prometheus
Web & APIs	Flask, FastAPI, Apache, Pydantic
Workflow	LangGraph
Machine Learning & AI	TensorFlow, PyTorch, OpenCV, DLib, Pandas
Visualization & UI	Dash Plotly, Qt/PyQt

[Operating System]

Linux	Proficient (daily use, primary os)
Windows	Proficient

[Electronics & Hardware]

Embedded boards	Raspberry Pi, NVIDIA Jetson (TX1/TX2, Xavier), AVR, Arduino
Sensors	**modules and sensors**, including cameras, temperature sensors, distance sensors etc.