

Hootan Seifi

AI Engineer · LLM & Agentic Systems

+98 935 775 9499

hootanseifi90@gmail.com

linkedin.com/in/hootan-seifi-42b180124

github.com/hnjgp

Tehran, Iran

SUMMARY

AI Engineer specializing in the **design and implementation of LLM-based systems**, from agent architecture and tool calling to RAG, vector search, fine-tuning, evaluation and guardrails. Focused on building **production-oriented** AI applications and **testable, secure and maintainable** systems. M.Sc. in Electrical Engineering (Artificial Intelligence), with a background in Machine Learning, Data Science and Data Engineering.

FEATURED PROJECT

Astra — Enterprise AI Agent Platform2026 – Present

Designed, architected and built independently, end-to-end

github.com/hnjgp/astra-ai-platform

Python · FastAPI · LangGraph · OpenAI · PostgreSQL + pgvector · Docker · Prometheus · GitHub Actions

407

automated tests

93%

test coverage

CI eval gate

Recall@k · MRR · routing

- What it is:** backend for an enterprise AI assistant. An LLM agent decides on its own whether it needs internal documents (RAG), live system tools, both, or neither, and every step runs inside security boundaries.
- Agent with LangGraph:** model → tools → model state machine with a recovery node, planning, and human-in-the-loop approval for sensitive tools via `interrupt()`, with a PostgreSQL checkpointer.
- Agentic RAG:** retrieval exposed as a tool the model calls when needed; overlapping chunking, `text-embedding-3-small`, pgvector search with an HNSW index, reranking and permission-aware retrieval.
- Tool layer and security:** centralized ToolExecutor with allow-list, RBAC, schema validation and retry; JWT auth, input/output guardrails, prompt-injection detection (English and Persian), secret redaction and rate limiting.
- Observability and MCP:** JSON logs with request ids, Prometheus metrics for model, embedding, tool and RAG calls, token usage and cost tracking; MCP server over stdio.
- Quality and delivery:** 407 unit, integration and end-to-end tests against real pgvector, 93% coverage; evaluation harness (Recall@k, MRR, route accuracy, groundedness) as a CI regression gate; Docker Compose and GitHub Actions.

CORE EXPERTISE

- LLM Application Engineering
- Agentic AI & Multi-Agent Systems
- RAG & Semantic Retrieval
- LLM Fine-tuning (QLoRA / PEFT)
- Retrieval & LLM Evaluation
- AI Guardrails & Security
- Production AI Backend (FastAPI)

TECH STACK

LLM & Agents

OpenAI / Anthropic APIsLangGraph

LangChainOpenAI Agents SDK

CrewAIAutoGenMCP

Hugging FaceQLoRA / PEFT

Tool CallingStructured Outputs

Vector Search & RAG

pgvectorChromaFAISS

EmbeddingsChunkingReranking

Recall@K

Backend & Deployment

PythonFastAPIPostgreSQL

DockerpytestGit

AWS (Bedrock, SageMaker, Lambda)

TerraformGitHub ActionsModal

Machine Learning

PyTorchscikit-learnXGBoost

NLPComputer Vision

Pandas / NumPy

Data Engineering

SparkKafkaAirflowSQL Server

ETL / ELT

CURRENTLY WORKING ON

Azmanegar — AI Clinical Decision Support Architecture design
LLM-based clinical decision support platform: lab result ingestion via HL7 / FHIR, LOINC / UCUM normalization, and an ML layer for anomaly detection, trend forecasting and dose-response analysis on patient data.

OTHER LLM & AGENTIC AI PROJECTS

- **QLoRA fine-tuning:** fine-tuned Llama 3.1 8B to predict product prices from descriptions, benchmarked against traditional ML baselines and fine-tuned frontier models (Hugging Face, Weights & Biases).
- **Autonomous multi-agent system:** deal discovery combining a fine-tuned model, RAG over Chroma and an ensemble model; serverless deployment on Modal.
- **Trading agents with MCP:** autonomous trader agents accessing market data, accounts and memory through MCP servers (OpenAI Agents SDK).
- **Deep research agent & multi-agent teams:** parallel web research with OpenAI Agents SDK, a browser agent with LangGraph and Playwright, and role-based teams with CrewAI and AutoGen.
- **Production deployment on AWS:** LLM and agent services on Bedrock, SageMaker and Lambda; infrastructure as code with Terraform and CI/CD with GitHub Actions.

EXPERIENCE

Senior Solar Panel Maintenance Specialist 2025
[Taban Energy Power Plant \(Pasargad Bank\)](#)

Fundamental Analyst & Trader, Financial Markets 2022 – 2024
[Novin Eghtesad Technologies](#)

- Fundamental analysis and trading in financial and crypto markets; evaluated blockchain projects and developed a trading bot in Python.

Technical Specialist, App Review & Development 2021 – 2022
[Cafe Bazaar](#)

R&D Specialist 2019 – 2020
[MAPNA Electric & Control Engineering \(MECO\)](#)

MACHINE LEARNING PROJECTS

- **Fraud detection** on Ethereum transactions with imbalanced data using XGBoost.
- **Sentiment-driven forecasting:** impact of tweets on Nvidia stock price using NLP and time series.
- **Computer vision:** traffic sign classification (GTSTRB) with CNNs, and a dangerous-overtaking warning system based on image processing.

EDUCATION

M.Sc. Electrical Engineering
Artificial Intelligence
Amirkabir University (Tafresh Branch)
2016 – 2018

B.Sc. Electrical Engineering
Electronics, Integrated Circuits
Amirkabir University (Tafresh Branch)
2011 – 2015

COURSES & CERTIFICATES

LLM Engineering: Master AI, Large Language Models & Agents
Ed Donner · Udemy

Machine Learning Specialization
Andrew Ng · DeepLearning.AI & Stanford

Financial Trading in Python
DataCamp · 2025
datacamp.com/completed/.../cc3de6b4

LANGUAGES

English (Advanced)
Persian (Native)

INTERESTS

- Blockchain projects
- Financial markets & algorithmic trading
- AI for financial markets

5 years of experience in fundamental analysis, trading and blockchain projects