NAVID HEJAZI

AI Engineer | RAG, Agentic Systems & Cloud

London, UK | 07436436342 | hejazi_navid@yahoo.com | LinkedIn | GitHub | Portfolio

PROFILE

Software engineer with commercial full-stack development and cloud computing experience, focused on building applied AI systems through hands-on agentic and RAG projects. Experienced in Python, Flask, AWS, multi-agent orchestration, hybrid retrieval, and LLM evaluation. Brings a strong engineering foundation to AI development, with an emphasis on measurable results, reliable architecture, and real-world delivery.

TECHNICAL SKILLS

AI / LLM Engineering: RAG (LangChain, hybrid search, ChromaDB, Pinecone, BM25), agentic systems (CrewAI, OpenAI Agents SDK, MCP), LLM evaluation and observability (LangSmith, tracing), prompt engineering, Voice AI (ElevenLabs, Twilio)

Software Engineering: Python, Flask, JavaScript, React, Node.js, SQL, RESTful APIs, WebSockets, Git

Cloud, Automation and DevOps: AWS (EC2, S3, RDS, Lambda, SageMaker, ECS Fargate, Bedrock, SQS, DynamoDB), Docker, CI/CD, GitHub Actions, n8n

SELECTED AI PROJECTS

Agentic Software Engineering Platform | Portfolio | GitHub

- Designed a cloud-deployed CrewAI platform that converts a natural-language requirement into a working Flask application through seven specialist agents covering planning, architecture, development, testing, debugging, and review.
- Architected an asynchronous AWS pipeline using ECS Fargate, SQS, DynamoDB, and S3 to separate long-running agent jobs from the web interface, with persistent state, live progress tracking, and automatic retries.
- Implemented controlled file I/O and pytest execution tools, enabling agents to generate tests, diagnose failures, and iteratively improve their code.
- Validated the workflow end-to-end by generating a functional application with requirements, architecture, testing, and review documentation from a single requirement.

Advanced RAG System with Evaluation and Optimisation | Portfolio | GitHub

- Built and compared two RAG pipelines: a vector-search baseline and a hybrid retrieval system designed to test whether targeted retrieval improvements outperform standard vector search.
- Combined semantic chunking, query rewriting, multi-query retrieval, ChromaDB vector search, and BM25 keyword search, followed by deduplication and LLM reranking.
- Created an evaluation framework measuring retrieval with MRR, nDCG, and coverage, and generation quality with accuracy, completeness, and relevance.
- Improved MRR by 10.7% and the answer-accuracy score by 14.6% relative to the baseline pipeline.

PROFESSIONAL EXPERIENCE

Full-Stack Developer | Tensator Group, Milton Keynes, UK — 02/2024 - Present

- Delivered two customer-facing queue management products with cross-functional teams, integrating real-time data to improve operational efficiency.
- Developed backend solutions using Python, Flask, and SQL, with WebSocket communication for real-time data updates, system configuration, and improved data accessibility.
- Created SQL queries and Chart.js dashboards to visualise operator availability, workload, and efficiency metrics for retail and airport environments.
- Integrated smart camera APIs to automate customer guidance and support real-time decisions based on the number of people in each area.

Junior Full-Stack Developer | Konica Minolta, London, UK — 02/2022 - 07/2022

- Developed Flask and Python web applications with an emphasis on clean code and unit testing.
- Designed and tested sign-in pages, applying input validation and unit tests to protect data integrity.
- Collaborated in an agile team using Bitbucket for version control and peer code reviews.

EDUCATION

Master's Degree in Computer Software Engineering - Azad University

Graduated with Distinction

- **Core Modules:** Advanced Programming, Data Structures, Machine Learning, Software Engineering
- **Thesis: Activity Recognition in Smart Homes Using Support Vector Machines (SVM),** applied SVM with tailored preprocessing on a public smart home dataset to improve activity recognition accuracy for dementia care.

Full-Stack Web Development Bootcamp - Code Your Future, London

Full-time | Aug 2022 - Dec 2023

- Built full-stack web applications using JavaScript, React, Node.js, and SQL, applying MVP principles to deliver functional solutions under tight deadlines.
- Collaborated in Agile teams using Slack and Miro for project planning, leading daily stand-ups and retrospectives.
- Deployed scalable applications using Docker Compose, CI/CD pipelines (GitHub Actions), and AWS services (S3, RDS, EC2) for automated cloud deployments.

PROFESSIONAL DEVELOPMENT

AI Engineer Production Track: Deploy LLMs & Agents at Scale - Udemy (2026)

- Designed and deployed serverless AI cloud architectures using Lambda, S3, SQS, CloudFront, and API Gateway, integrated with Amazon Bedrock and SageMaker.
- Practised infrastructure-as-code and CI/CD workflows using Terraform and GitHub Actions.
- Built and secured multi-agent SaaS applications with the OpenAI Agents SDK, including production observability and guardrails.

AI Engineer Core Track: LLM Engineering, RAG, QLoRA, Agents - Udemy (2026)

- Built LLM-powered applications using tool calling, RAG pipelines, embeddings, vector retrieval, and LoRA/QLoRA fine-tuning with Hugging Face.
- Evaluated and compared frontier and open-source models on classification and price-prediction tasks.
- Applied agentic workflows and multi-agent system design within broader LLM engineering projects.

AI Engineer Agentic Track: The Complete Agent & MCP Course - Udemy (2026)

- Built AI agents with OpenAI Agents SDK, LangGraph, CrewAI, AutoGen, and MCP, including state management and external tool integration.
- Practised multi-agent orchestration patterns including sequential pipelines, hierarchical delegation, and state-driven workflows.
- Implemented task delegation and external tools such as web search and content generation across agentic frameworks.