Sepehr Rezaee

sepehrrezaee2002@gmail.com | github.com/SepehrRezaee | linkedin.com/in/sepehr-rezaee | sepehrrezaee.com

Agentic AI Engineer with 5+ years of hands-on experience in designing, orchestrating, and deploying multi-agent AI/LLM systems for production. Expert in Python, Kubernetes, Docker, and LLM-based agent platforms (LangChain, RAG, OpenAI GPT, Anthropic Claude). Track record of building scalable, autonomous, business-driven AI agents and pipelines for SaaS and enterprise, with a focus on agent communication, memory, orchestration, and outcome optimization.

Core Skills & Technologies

- **AI Agent Orchestration:** LangChain, LangGraph, LlamaIndex, Multi-Agent Systems, RAG, SPAR (Sense, Plan, Act, Reflect), Prompt Engineering
- **Programming:** Python (expert), C++, Java, C#
- **Infrastructure:** Docker (expert), Kubernetes (expert), AWS (SageMaker, EC2), GCP, Azure
- **Databases:** Pinecone, Weaviate, Chroma, PostgreSQL (pgvector), MongoDB, Redis
- **LLMs:** OpenAI GPT (3/4), Anthropic Claude, Google Gemini, Hugging Face Transformers
- **MLOps:** MLflow, Airflow, Celery, Prometheus, Grafana, ELK, FastAPI, Flask, REST/GraphQL APIs
- **Other:** Knowledge Graphs (Neo4j), Multimodal AI, Speech/Text Interfaces, SaaS architecture

Professional Experience

AI Engineer, Agentic Systems *PropTy Global, Remote*
Aug 2024 – Present

- Architected and deployed production-ready multi-agent LLM systems (LangChain, custom RAG), driving autonomous recommendation and business decision workflows (85%+ completion).
- Developed robust agent-to-agent protocols and memory modules for context-aware, goal-driven agents.
- Integrated Docker/Kubernetes for scalable, low-latency deployments (sub-100ms API), and implemented advanced monitoring (Prometheus, Grafana).
- Connected agent actions to live business KPIs, building feedback/evaluation loops for agent optimization.
- Collaborated on outcome tracking and continuous agent improvement pipelines.

Chief AI Officer & Multi-Agent Architect *Novel Mind Scientist, Tehran, Iran*
Oct 2022 – Present

- Led the full-stack delivery of LLM-powered agents for SaaS, healthcare, and education, integrating vision, text, and knowledge graph data.
- Implemented multi-agent orchestration (LangChain, Celery) and business process automation pipelines.
- Provided technical leadership: code reviews, design standards, agent evaluation, documentation, and knowledge transfer.

Research Assistant – Agentic AI & Security *Sharif University & Shahid Beheshti University*
2023 – 2025

- Developed advanced agentic ML pipelines for secure, robust AI—including RAG systems, agent routing/hand-off, and memory management.
- Published/Submitted papers to NeurIPS, ICCV on autonomous agent security, evaluation, and optimization.
- Mentored junior engineers and contributed to open-source agentic AI codebases.

Project Manager, Agentic ML SaaS *NovaVira, Tehran, Iran*
Mar 2023 – Feb 2024

- Delivered a Django-based agentic recommender platform (LangChain, GCP, Docker) with hybrid search and automated workflow.
- Oversaw Agile/CI-CD, ensuring reliability and fast iteration on agent architectures.

Education

B.Sc. in Computer Science *Shahid Beheshti University, Tehran*
2021–2025 (expected), GPA: 3.4/4.0

Selected Agentic AI Projects

- **Multi-Agent RAG Platform:** Designed, orchestrated, and deployed a scalable agentic system for business process automation, leveraging LangChain, custom agent protocols, and Pinecone/Weaviate vector DBs (2024).
- **Agent Memory & Routing:** Built and productionized agent memory/long-term context systems, enabling dynamic agent routing, escalation, and autonomous hand-off.
- **LLM Evaluation & Optimization:** Created automated evaluation and feedback pipelines for agentic LLM workflows, tracking business outcomes and supporting continuous improvement.
- **AI Model Security:** Developed adversarially robust agentic ML pipelines, published in NeurIPS 2024.

Selected Publications

- DISTIL: Data-Free Inversion of Suspicious Trojan Inputs via Latent Diffusion. (ICCV 2025, submitted)
- Scanning Trojaned Models Using Out-of-Distribution Samples. (NeurIPS 2024, accepted)
- Comparison of Pre-Training and Classification Models for Early Detection of Alzheimer's Disease Using MRI. (I4C 2023)

Awards

- Best Ideator Award, National Young Scientists Festival (2023)
- Placed 352nd of 150,000 in National Entrance Exam (2020)

Languages

Persian (Native), English (Professional)

References

Prof. Kourosh Parand - k_parand@sbu.ac.ir
Prof. Mohammad Hossein Rohban - rohban@sharif.edu
Prof. Mohammad Sabokrou -mohammad.sabokrou@oist.jp
Prof. Mathis Mackenzie mackenzie.mathis@epfl.ch