Sepehr Rezaee

AI Architecture & Multi-Agent Systems — RAG Platforms & Safety — MLOps at Scale
sepehrrezaee2002@gmail.com — github.com/SepehrRezaee — linkedin.com/in/sepehr-rezaee — sepehrrezaee.com

Summary

Senior AI architect specializing in **agentic LLM systems**, **retrieval-augmented generation**, and **safety-first** platforms. 5+ years designing, evaluating, and operating production **multi-agent** services with **SLOs**, **error budgets**, and **cost/latency** controls. Expert in **Python**, **Docker/Kubernetes**, and **LangChain/LangGraph/LlamaIndex**. Known for **architecture governance**, **standardization**, and **cross-functional leadership** across SaaS and enterprise workloads; research grounding with peer-reviewed publications (ICCV/NeurIPS).

Leadership & Architecture Highlights

- **Platformization:** Productized a reusable *multi-agent RAG platform* with standardized *prompts, tools, memory, eval harnesses*, and *guardrails*, enabling rapid service rollout and consistent reliability.
- **Reliability & Observability:** Drove *SLOs, error budgets, tracing, dashboards* (Prometheus/Grafana/ELK) and data/feature pipelines for stable throughput during peak load.
- **Safety & Governance:** Established *prompt/tooling governance*, red-team evaluation, fallback policies, and incident response patterns aligned to enterprise risk tolerance.
- **Orchestration Excellence:** Designed agent-to-agent protocols, *long-term memory*, and dynamic *routing/hand-off* for autonomous workflows; optimized inference via *vector caching*, *retrieval tuning*, and *model routing*.
- **Research to Production:** Translated *ICCV/NeurIPS* research in model security and evaluation into hardened production practices.

Core Skills & Technologies

- **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/Vector:** Pinecone, Weaviate, Chroma, PostgreSQL (pgvector), MongoDB, Redis
- **LLMs:** OpenAI GPT-3/4, Anthropic Claude, Google Gemini, HF Transformers
- **MLOps & Services:** MLflow, Airflow, Celery, Prometheus, Grafana, ELK, FastAPI, Flask, REST/GraphQL
- **Other:** Knowledge Graphs (Neo4j), Multimodal AI, Speech/Text Interfaces, SaaS Architecture

Professional Experience

Senior LLM Engineer *AIR Property, Dubai*
Aug 2025 – Present

- Architected and shipped **production LLM services** (RAG + agents), owning *model selection*, *agent/prompt design*, *eval harnesses*, and *fallback trees*; improved answer quality with strict **latency/cost** budgets.
- Introduced an **architecture blueprint** for agent services and a *governed toolchain* (tool registries, policies, approvals), raising reliability and reusability across projects.
- Built modular **APIs** with tests and **dashboards** for *SLOs, error budgets, safety metrics*; partnered with product/security to align thresholds and incident response.
- Drove **data curation**, **vector caching**, and **inference optimization** to stabilize throughput under peak load; led capacity planning and rollouts on Kubernetes.

AI Engineer, Agentic Systems *PropTy Global, Dubai*
Aug 2024 – Sep 2025

- **Architected multi-agent systems** (LangChain + custom RAG) for autonomous recommendations/decisions; achieved **85%+** end-to-end task completion in business workflows.
- Implemented *agent-to-agent protocols* and *memory* for context-aware planning and goal execution; standardized *routing/hand-off* patterns.

- Productionized on **Docker/Kubernetes** with sub-100ms API path for critical endpoints; instrumented with **Prometheus/Grafana** and centralized logging.
- Closed the loop to live **KPIs** with automated evaluation/feedback for continuous improvement and drift monitoring.

Chief AI Officer & Multi-Agent Architect *Novel Mind Scientist, Tehran, Iran*
Oct 2022 – Sep 2025

- Led delivery of **LLM-powered agents** across SaaS/health/education, integrating **text/vision** and **knowledge graphs** with measurable **SLAs**.
- Scaled **multi-agent orchestration** (LangChain, Celery) and automated business processes; established *design reviews*, *eval protocols*, and *knowledge transfer*.
- Drove **engineering standards** (docs, ADRs, onboarding guides) to accelerate adoption and reduce operational risk.

Research Intern, Safe & Generative AI *Mathis Lab, EPFL (Switzerland)*
May 2025 – Sep 2025

- Co-authored *ICCV 2025 (accepted)* paper on **DISTIL**: data-free diffusion-based trigger inversion for Trojaned models; new SOTA on BackdoorBench (+7.1% acc) and object-detection scanning (+9.4%).
- Built **latent-diffusion** pipelines with classifier-guided feedback to expose adversarial vulnerabilities for safer AI.
- Developed **zero-shot, data-free defenses** and ran large-scale benchmarks; published best practices for robust evaluation.

Research Assistant — Agentic AI & Security *Sharif Univ. & Shahid Beheshti Univ.*
2023 – 2025

- Prototyped **secure agentic ML pipelines**: RAG, routing/hand-off, memory management.
- Published/submitted work to **NeurIPS/ICCV** on agent security, evaluation, and optimization; mentored junior engineers.

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

- Delivered Django-based **agentic recommender** (LangChain, GCP, Docker) with hybrid search and automated workflows.
- Ran **Agile CI/CD** to accelerate iteration on agent architectures and platform reliability.

Education

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

Architecture & Platform Work

- **Multi-Agent RAG Platform (2024):** Orchestrated scalable automation with LangChain, custom protocols, Pinecone/Weaviate; standardized *evals, guardrails, fallback policies*.
- **Agent Memory & Routing:** Long-term memory + dynamic routing and escalation; autonomous hand-off between tools/agents; reusable templates for new use-cases.
- **LLM Evaluation & Optimization:** Automated eval + feedback tied to business outcomes; model/routing comparisons and prompt A/B to control cost/latency.
- **AI Model Security:** Adversarially robust pipelines; **NeurIPS 2024 (accepted)**; safety red teaming integrated into CI.

Selected Publications

- **DISTIL:** Data-Free Inversion of Suspicious Trojan Inputs via Latent Diffusion. *ICCV 2025* (**accepted**).

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

Languages

Persian (Native); English (Professional)

References

Available upon request (*Prof. M. H. Rohban; Prof. Mackenzie Mathis; Prof. M. Sabokrou; Prof. K. Parand*)