

SARA SADR JAHANI

AI / NLP Engineer — GenAI & RAG Systems

sarasadrjahani@gmail.com | +98 912 599 2674 | [linkedin.com/in/sadrjahani](https://www.linkedin.com/in/sadrjahani) | github.com/sadrjahani

SUMMARY

- AI/NLP Engineer with 7 years in Natural Language Processing (NLP) and transformer models (research since 2019, SBERT / Siamese BERT) and approximately 2 years building Large Language Model (LLM) applications: Retrieval-Augmented Generation (RAG) pipelines, AI agents (Model Context Protocol, MCP) and Neo4j-backed retrieval for industrial and fintech use cases.
- Cut prompt-design iteration time by ~15% with an automated prompt-optimization workflow and improved semantic-search retrieval by ~30% with a Neo4j-based RAG system.
- Built a fully local PDF RAG pipeline (FAISS, Ollama) with a 68-question retrieval evaluation; building public projects on ReAct agents, LangChain and GraphRAG (github.com/sadrjahani).
- English (IELTS 6.5), German B1 (conversational; B2 course in progress), Persian (native).

TECHNICAL SKILLS

LLM / GenAI: Retrieval-Augmented Generation (RAG), LangChain, Model Context Protocol (MCP) Server, AI agent design, prompt engineering and optimization, GPT, LLaMA, LoRA/PEFT (in progress)

Retrieval & Data: FAISS, SentenceTransformers, Neo4j, vector retrieval, PyMuPDF, SQL, web scraping

Engineering: Python, FastAPI, Docker (working knowledge), Google Cloud Platform (GCP), Linux, Git, Ollama

NLP / ML: PyTorch, scikit-learn, Transformers (BERT, SBERT, Augmented SBERT), LSTM, XGBoost, time-series forecasting, PySpark

PROFESSIONAL EXPERIENCE

Independent Projects & Upskilling

Sep 2026 – Present

- Built and evaluated a local PDF RAG system (see Selected Projects); building a ReAct agent from scratch, a LangChain rebuild and a GraphRAG project in public repositories.

Machine Learning Engineer

Mar 2025 – Aug 2026

Freelance

- Engineered RAG pipelines with LangChain and served them via API endpoints for industrial clients, cutting design-iteration time by ~15% through automated prompt optimization.
- Built AI agent systems on MCP Server to standardize how LLMs access enterprise data sources.
- Designed reproducible end-to-end AI pipelines, from predictive-model optimization through deployment, packaging select services with FastAPI and Docker.

AI Engineer

Sep 2024 – Feb 2025

Goodfolio, London (Remote)

- Built consent-based data-collection pipelines (Google Forms/Workspace to GCP) where users voluntarily shared professional profile data (e.g., from LinkedIn) with explicit permission; applied LLMs to analyze it and power end-to-end service delivery.
- Architected Neo4j graph databases for semantic search in LLM applications, improving relevant data-insight retrieval by ~30%.
- Raised LLM output accuracy by ~15% with few-shot and structured prompting, working alongside a UK-based product team.

Machine Learning Specialist

Nov 2022 – Jul 2024

Freelance

- Delivered fraud-detection and sentiment-analysis models for unstructured text, reaching 95% precision.
- Built and validated time-series forecasting frameworks (LSTM, XGBoost) with rigorous testing.
- Developed Django-based APIs that improved pricing efficiency by ~30%, plus web-scraping pipelines for market intelligence.
- Shipped multilingual (English/Persian) NLP solutions that cut manual processing time by ~40%.

Data Science Specialist

Jul 2021 – Nov 2022

Sabke Zandegi Khallagh (Alinland e-commerce platform), Tehran

- Built automated web-scraping and data pipelines that kept e-commerce datasets current.
- Analyzed e-commerce data trends to optimize marketing strategies and pricing decisions.

Research Assistant <i>Intelligent Systems Group, Institute for Advanced Studies in Basic Sciences (IASBS), Zanjan</i>	2018 – 2022
- Designed Siamese BERT architectures for semantic scoring of augmented sentences, benchmarked against SBERT and Augmented SBERT baselines, improving Semantic Textual Similarity (STS) performance. - Researched LSTM/XGBoost-based traffic-flow prediction.	
Early Career — IT & Network Administration <i>Alzahra University; Arian Saeed Industrial Group (ASIGI), Tehran</i>	2016 – 2018
Supported virtualization servers and network services; built an internal management dashboard for decision-making.	

SELECTED PROJECTS

Local PDF RAG System Python, FAISS, SentenceTransformers, Ollama github.com/sadrjahani/rag-ga-system	
- Built a fully local RAG pipeline for question answering over PDF books: PDF extraction (PyMuPDF), section-aware chunking with overlap, embeddings (all-MiniLM-L6-v2), and exact FAISS cosine-similarity retrieval. - Implemented grounded answer generation with a local Ollama model (phi4-mini), constrained to retrieved context with source citations. - Designed a modular architecture with index/document alignment and embedding-dimension integrity checks. - Evaluated retrieval on a 68-question golden set: increasing top-k from 3 to 5 raised context recall from 0.73 to 0.83.	
ReAct Agent From Scratch (Python, uv, Ollama) github.com/sadrjahani/react-agent-from-scratch In progress	
- Building a ReAct (Reasoning + Acting) agent in plain Python without agent frameworks. - Implemented a provider-agnostic LLM wrapper supporting Anthropic, OpenAI and local Ollama models, configured via environment variables.	

EDUCATION

M.Sc. Computer Science — Data Mining <i>Institute for Advanced Studies in Basic Sciences (IASBS), Zanjan GPA 17.38/20</i>	2018 – 2022
Thesis: Scoring Augmented Sentences Using Autoencoder and Semantic Textual Similarity (STS) with Siamese BERT	
B.Sc. Computer Science <i>Alzahra University, Tehran</i>	2012 – 2016

CERTIFICATIONS

Neural Networks and Deep Learning | AI Agents Fundamentals

LEADERSHIP & AWARDS

- President, Computer Science Association, Alzahra University (2012 – 2016): mentored 50+ students through career-roadmap workshops. - Medalist, International Inventions Competition, Kuwait (2011); Best Student Researcher, Zakaria Razi Research Institute (2009).	
---	--