Masih Moafi

AI Engineer - Agentic Systems, RAG, and MCP

Isfahan, Iran
masihmoafi12@gmail.com | masihmoafi.com
github.com/MasihMoafi | linkedin.com/in/MasihMoafi1
Open to international relocation and remote roles

Applied AI Engineer building Generative AI and Large Language Model (LLM) systems, agentic workflows, Retrieval-Augmented Generation (RAG), and production Python backends. Experience spans Model Context Protocol (MCP), tool calling, structured outputs, context and memory management, hybrid retrieval, evaluation, and infrastructure from API design through Docker and CI/CD.

> **EVIDENCE** Marketplace backend aggregating approximately 20,000 listings | 47-65% lower peak agent context across three paired workloads | 90.9% full retrieval on a mixed-code evaluation corpus

CORE EXPERTISE

Agents and LLM systems: Multi-agent orchestration, tool and function calling, tool schemas, MCP, LangGraph, structured outputs, LLM APIs, prompt and context engineering, memory and state management

Retrieval engineering: Retrieval-Augmented Generation (RAG), BM25, vector search, hybrid retrieval, Reciprocal Rank Fusion (RRF), cross-encoder reranking, retrieval evaluation, LanceDB, Qdrant

Backend and infrastructure: Python, FastAPI, Django, Pydantic, REST APIs, PostgreSQL, Redis, Celery, Docker, Nginx, GitHub Actions, Linux, Bash

Evaluation and machine learning: Behavioral evals, evaluation frameworks, regression testing, token accounting, PyTorch, Transformers, LightGBM, CatBoost, Ollama, CUDA

PROFESSIONAL EXPERIENCE

AI/ML Engineer - AI Can Do

May 2025 - Aug 2026

Backend engineering, LLM applications, and DevOps

Muscat, Oman

- Built the backend for Boomdeal, a C2C car-search marketplace aggregating approximately 20,000 listings, including scraping and ingestion infrastructure.
- Developed a conversational search workflow that translated natural-language vehicle requests into structured database queries through OpenRouter LLM APIs.
- Shipped Django, PostgreSQL, Redis, and Celery services with validation and error contracts; configured Docker, Nginx, SSL, and GitHub Actions CI/CD.

Python | Django | PostgreSQL | Redis | Celery | Docker | Nginx | OpenRouter API | CI/CD

AI/ML Engineer - Pejvak Rayane

Sept 2024 - Dec 2025

Agent tooling, hybrid retrieval, and structured outputs

Isfahan, Iran

- Built Model Context Protocol servers and typed tools that exposed local system and utility capabilities to AI agents.
- Developed hybrid retrieval combining semantic vector search, BM25 lexical search, and open-source model-based reranking.
- Implemented scoped memory and state, structured-output validation, and agent orchestration with LangGraph and Pydantic.

Python | MCP | LangGraph | Pydantic | Qdrant | BM25 | RAG | Reranking

ML Engineer (Freelance) - Financial Market Analysis

Sept 2024 - Apr 2025

Modeling, evaluation, and MLOps

Tehran, Iran

- Built an end-to-end EUR/USD data, training, prediction, and backtesting pipeline using five years of one-minute market data.
- Developed a dual-tower Gated Transformer Network and LightGBM/CatBoost baselines with class-imbalance objectives and temporal evaluation.
- Vectorized Pandas and NumPy feature computation, managed PyTorch runs on Vast.ai A100/H100 GPUs, and integrated predictions with MetaTrader 5.

PyTorch | LightGBM | CatBoost | Pandas | NumPy | Vast.ai | MetaTrader 5 | MQL5

SELECTED ENGINEERING EVIDENCE

Elpis - Context and Continuity for Coding Agents GitHub | Project report

Open-source Rust environment built on OpenAI Codex for inspectable context, durable continuity, and provider-neutral agent operation

- Built context admission controls, memory and continuity state, tool-output filtering, selective post-turn pruning, durable transcripts, and an inspectable Context Ledger.
- Added deterministic work-graph validation for dependencies, write boundaries, and evidence-gated verification around delegated agent tasks.
- Across three paired, byte-identical workloads using the same model and source commit, Elpis used 47-65% less peak context per request; this result measures context footprint, not task quality.

Rust | Ratatui | Agent harnesses | Context engineering | MCP | CI/CD | Evaluation

rag-mcp - Local Hybrid Retrieval for Coding Agents GitHub | Live demo | Report

Read-only MCP retrieval service for source-grounded repository and document search

- Built an end-to-end retrieval pipeline with document ingestion, Tantivy BM25, LanceDB vector search, RRF fusion, cross-encoder reranking, and exact source paths.
- Published evaluations across five real corpora; one mixed Rust, Python, and notebook corpus reached 90.9% full retrieval and 100% at-least-partial retrieval.

Python | MCP | LanceDB | Tantivy BM25 | RRF | Cross-Encoder | Apache Arrow | Retrieval evaluation

Multivector Safety Analysis of gpt-oss:20b DOI | GitHub | Report

Controlled LLM safety evaluation of prompt framing, dialogue state, tool use, and reasoning-channel exposure

- Designed fixed-prompt, seed-based behavioral evaluations with mechanically checked endpoints and explicit limits on every claim.
- Recorded 78/100 versus 97/100 exact tool calls under matched framing conditions; a protected string appeared in reasoning in 25/30 trials and visible answers in 13/30.

Python | LLM evaluation | Red-teaming | Tool calling | Structured outputs | Ollama | Kaggle GPUs

ADDITIONAL ENGINEERING WORK

WhisperType: Built local Linux/NVIDIA voice input with whisper.cpp, CUDA transcription, hotkey capture, direct cursor insertion, and optional Gemini text refinement. github.com/MasihMoafi/WhisperType

Transformer implementations: Implemented neural networks, RNNs, LSTMs, WaveNet, tokenization, Transformers, GPT-2, and RoPE from scratch in Python and PyTorch; documented the work at masihmoafi.com/blog.

EDUCATION AND LANGUAGE

M.Sc. in Artificial Intelligence and Robotics 2025 - 2027 (expected)
Isfahan University of Technology (IUT), Isfahan, Iran

B.Sc. in Interior Architecture 2017 - 2022
Shiraz University of Arts, Shiraz, Iran

TOEFL iBT: 112/120 August 2024
Reading 30 | Listening 29 | Speaking 26 | Writing 27