



CONTACT

- sepehrradmard.ir
- sepehr80@gmail.com
- github.com/sepehr071
- linkedin.com/in/sepehr-radmard
- Shiraz, Iran

LANGUAGES

Persian	Native
English	Advanced

CORE SKILLS

STACK
Python · TypeScript · SQL · FastAPI · Next.js

LLM
OpenRouter · RAG · agents · Ollama

VOICE
LiveKit · Vapi · ElevenLabs · OpenAI Realtime

INFRA
Docker · GitLab CI · Keycloak · PostgreSQL

Sepehr Radmard

AI Engineer — LLM products & voice

Shiraz, Iran · sepehr80@gmail.com · github.com/sepehr071 · LinkedIn · sepehrradmard.ir

Professional Summary

AI Engineer building production LLM systems across RAG, voice agents, OCR, and AI infrastructure. I deliver end-to-end AI products from APIs and authentication to deployment, evaluation, and observability. Previously AI Developer at Danak; nine years of freelance full-stack web development before that.

Technical Skills

Languages	Python, TypeScript, SQL
Frameworks	FastAPI, Next.js
LLM Stack	OpenRouter, Ollama, Agents, RAG, LangChain/LlamaIndex, DLP · Gemini 3.6 / Claude Opus 5 / GPT-5.6
Vector Search	Pinecone, pgvector, Cohere Rerank, LlamaCloud, hybrid dense+sparse, multi-query fusion, RRF
Voice AI	LiveKit, Vapi, ElevenLabs, OpenAI Realtime, in-call tool calling, barge-in
Infrastructure	Docker, GitLab CI/CD, Nginx, Linux, Keycloak, PostgreSQL, Unsloth LoRA / QLoRA, 4-bit quant
Observability	Langfuse, DeepEval, Sentry, Datadog

Selected Projects

Novis AI aggregator

One login for chat, agents, and document tools that must pick a model by quality, latency, and cost.

- Solo full-stack: FastAPI, UI, Keycloak SSO, Docker/CI. Every LLM call goes through **OpenRouter**; the user or a router picks Gemini / Claude / GPT. Streaming live chat on several Uvicorn/Gunicorn workers behind Nginx.
- Per-org token budgets. DLP before the prompt leaves: regex on API keys, Iranian national ID, cards, phones, emails, and uploads, plus a local **Ollama** LLM that blocks confidential text.
- Designed to support **130+ concurrent users** in production (stress-tested past that without a crash). Web + Android. Open-source sibling on GitHub (polymind).

Tariq — religious RAG

Persian + Arabic chatbot over **50+ books**, 800+ pages each. Dirty scans, mixed script, and answers grounded in the indexed corpus rather than unrestricted LLM generation.

- I owned ingest solo: scan/OCR → clean → chunk → hybrid dense+sparse indexes. Multi-query fusion (LLM paraphrases + Reciprocal Rank Fusion), then one global Cohere rerank.
- Returns grounded answers only and refuses unsupported queries when the corpus lacks evidence. Same FastAPI + Docker + Gunicorn service later for law and Sepidar. LlamaIndex + LlamaCloud.



THIS PAGE

VOICE

LiveKit Agents · barge-in · in-call tools

ALSO

Meeting · ticketing · OCR

CONTACT

✉ sepehr80@gmail.com

🐙 github.com/sepehr071

Selected Projects

02 / 04

Realtime voice + telephony

Production-grade realtime voice agents with barge-in, tool calling, retrieval before response generation, and WebRTC telephony.

- One **LiveKit Agents** core for **EngageIQ** (EuroShop booth: QR → talk → tools capture the lead), **Beauty Lounge**, and **MorphoDerm**. In-call tools + ElevenLabs TTS. Shared engine; client-specific tools and prompts.
- **DFKI** is the fork: every avatar answer hits vector RAG first, then TTS, so the lab avatar does not invent. Webhooks for call events.

Meeting assistant

Overlapping Farsi speech. Delivers structured meeting minutes, decisions, action items, and follow-up emails from multilingual meeting recordings.

- Upload, microphone, or tab capture (Meet / Zoom / Teams). ElevenLabs Scribe produces captions and a speaker-labeled transcript.
- An LLM writes the pack: summary, decisions, action items, follow-up email. Keycloak login.

IT ticketing

Staff type a Farsi IT problem. Most should stop at the FAQ. Unsafe or unknown ones must hit a human.

- FAQ retrieval first; LLM triage only on a miss. Escalates on policy. Manager dashboard for queue load and handle time.
- Reads screenshots, sets priority. Ticket creates immediately — AI processing runs asynchronously after ticket creation, reducing perceived wait time.

Document OCR invoices · Iranian forms

Finance drops a scan. Fields they can approve — or the file never leaves the building.

- Air-gap: PP-OCRv5 on GPU plus a local Ollama / Qwen extract when data stays in-box; cloud VLM only if the file can leave.
- Messy Iranian forms and sefte (layout, stamps, Farsi amounts). Schema: parties, line items, bboxes. Batch → Excel. Human-in-the-loop review UI.



THIS PAGE

FEED0

ERP QA → GitLab write-up

LAB

Unsloth · Persian STT eval

LANGUAGES

Persian	Native
English	Advanced

CONTACT

- 🌐 sepehrradmard.ir
- ✉ sepehr80@gmail.com
- 🐙 github.com/sepehr071
- 📍 Shiraz, Iran

Selected Projects

Feedo

- Non-tech ERP ops describe a bug in Farsi. QA needs a GitLab-ready write-up — not an IT helpdesk ticket.
- Vision on the screenshot. Classified ticket plus paste-ready GitLab text. Manager approve / reject.
 - Next.js + FastAPI + Keycloak roles. Designed for non-technical staff who report issues without writing structured bug reports.

LLM compression — Unsloth

- Fine-tune and shrink models so they run on one GPU instead of a hosted API.
- **Lab:** LoRA / QLoRA on 4-bit bases, distill a teacher into a student, measure the task not the loss curve, then GGUF.
 - **Shipped locally:** Ollama for Novis DLP (confidential-text block) and OCR field extract (Qwen) when files cannot leave.

Persian STT eval lab

Whisper (FA) vs NeMo FastConformer on an RTX 4060 — pick the Farsi ear before meeting/voice prod. Meeting STT in prod is still ElevenLabs Scribe.

How I ship

Full-stack 0 → MVP

- Blank repo to a URL with login. AI models are integrated as components within reliable production workflows.
- Loop first: who types, what the model may see, when a human approves (OCR review, ticket escalate, DLP block).
 - Then FastAPI + Next.js + Keycloak + Postgres. Farsi screens are RTL from day one, not a patch.
 - Docker, GitLab CI, Nginx on Linux. Once live, accelerate development with Claude Code, Cursor, Hermes Agent, OpenClaw, and MCP.



THIS PAGE

EXPERIENCE

Zariyal · Danak · Ramzagah · WordPress

EDUCATION

Shiraz M.Sc. AI · Yazd B.Sc. CS

LANGUAGES

Persian	Native
English	Advanced

CONTACT

- sepehrradmard.ir
- sepehr80@gmail.com
- github.com/sepehr071
- Shiraz, Iran

Experience + Education

Professional Experience

AI Engineer Zariyal Holding · May 2026 — Present

AI product engineer for the holding: aggregator, RAG, meetings, helpdesk, OCR, and realtime voice — the projects above.

- Solo full-stack on Novis and Tariq. LiveKit Agents in production for voice clients; DFKI is the RAG-before-speak fork.
- Own end-to-end delivery across backend APIs, frontend UI, authentication, deployment, and production infrastructure.

AI Developer Danak · Aug 2025 — May 2026

Internal LLM features for a non-technical ops team — chat, document reading, automation. Not the Zariyal products.

- Wired models into daily workflows. Built production features adopted by internal operations teams.
- Same habit: UI + API + a model, then hand it to staff. Scope is internal ops, not a public SaaS.

SEO Manager Ramzagah

Organic search for the product site — calendar, on-page SEO, traffic that is not only ads.

- Owned the content calendar and on-page work so pages stay findable.

Freelance Full-stack Web Developer 2016 — 2024

Nine years shipping sites: themes, plugins, front and back. That is why the AI products have real UIs and deploys.

- Built and deployed client websites, custom WordPress plugins, and backend integrations across 9 years of freelance work.

Education

Shiraz University

2023 — present

M.Sc. Computer Engineering — Artificial Intelligence

Yazd University

2020 — 2023

B.Sc. Computer Science