



CONTACT

- sepehradmard.ir
- sepehr80@gmail.com
- +98 902 441 9033
- github.com/sepehr071
- linkedin.com/in/sepehr-radmard
- 1 Jan 2001
- Shiraz, Iran

LANGUAGES

Persian	Native
English	Advanced

CORE SKILLS

- AI / LLM
Python, FastAPI, OpenRouter, Ollama, agents, RAG, DLP
- VOICE
LiveKit · Vapi · ElevenLabs · in-call tools
- EVAL
Langfuse · DeepEval · Sentry · Datadog
- AI CODING
Vibe coding · OpenClaw · Hermes · Claude Code

Sepehr Radmard

AI Engineer — LLM products & voice

Shiraz, Iran · +98 902 441 9033 · sepehr80@gmail.com · github.com/sepehr071 · LinkedIn · sepehradmard.ir

Professional Summary

I ship live AI products for Iranian companies — workspace, RAG, meetings, helpdesk, OCR, and realtime voice / telephony agents. I build the whole path (API, UI, auth, Docker/CI) and vibe-code the rest with Claude Code, Cursor, Hermes, and MCP. Previously AI Developer at Danak; nine years of WordPress before that.

Technical Skills

AI / LLM	Python, FastAPI, Next.js, OpenRouter, Ollama, agents, DLP · Gemini 3.6 / Claude Opus 5 / GPT-5.6 · pick by quality, latency, cost
RAG / vectors	Hybrid dense+sparse, multi-query fusion, RRF, Cohere rerank, LlamaCloud, Pinecone, pgvector, embeddings
Voice / telephony	LiveKit Agents, Vapi, ElevenLabs, OpenAI Realtime, in-call tool calling, barge-in latency, webhooks, conversation QA
Eval / obs	Langfuse, DeepEval, Sentry, Datadog — traces, agent evals, production errors
DevOps / MLOps	Docker, GitLab CI/CD, Nginx, GPU serving, Unsloth LoRA / QLoRA, 4-bit quant, distillation, Ollama / GGUF
AI coding	vibe coding — Claude Code, Cursor, Grok, Hermes Agent, OpenClaw, MCP, skills / subagents

Selected Projects

Novis AI

Company AI workspace: one login for chat, agents, and document tools — SSO, usage budgets, and DLP.

- Sized for about **130 people online at once** on live chat (stress-tested past that without a crash).
- Shipped as a full product: web + Android, Docker/CI on Linux. Open-source sibling on GitHub (uni-chat).

Meeting assistant

A team records a Farsi meeting. They leave with minutes, actions, and a follow-up email — not a raw WAV.

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

IT ticketing

Staff type a problem in Farsi. FAQ answers first; if that fails, an LLM triages until a human takes over.

- Reads screenshots, sets priority, escalates on policy. Manager dashboard for queue load and handle time.
- Fast ticket create — the AI runs in the background so the user is not staring at a spinner.



THIS PAGE

RAG

Tariq — 50+ books · hybrid + rerank

ALSO

OCR · Unsloth · voice · Feedo

CONTACT

✉ sepehr80@gmail.com

🐙 github.com/sepehr071

Selected Projects

02 / 03

Tariq — religious RAG

Persian + Arabic chatbot over a library of **50+ books**, each 800+ pages. I owned the full ingest: scan/OCR → clean text → chunk → vector indexes.

- Production retrieval API (LlamaIndex + LlamaCloud): hybrid dense + sparse search, multi-query fusion (LLM paraphrases + Reciprocal Rank Fusion), then one global Cohere rerank.
- Same service pattern for law and Sepidar corpora. FastAPI, Docker, Unicorn. Answers come from the books, not a naked LLM.

Document OCR invoices · Iranian forms

Finance drops a scan. They get fields they can approve — or the file never leaves the building.

- PP-OCRv5 on GPU when data stays in-box; cloud VLM when it can leave. On-prem extract uses a quantized local LLM (Ollama / Qwen).
- Human-in-the-loop review. Invoice schema (parties, line items, bboxes). Batch sefte → Excel.

Realtime voice + telephony

Speech-to-speech agents with RAG and in-call tools — not a toy WebRTC demo. Latency, barge-in, and call quality are the job.

- **EngageIQ** — EuroShop booth: QR → talk → tools capture the lead. LiveKit / Vapi-style realtime, ElevenLabs TTS.
- **DFKI** — lab avatar; every answer hits vector RAG first, then tools. Webhooks for call events.
- **Beauty Lounge / Morphoderm** — treatment desk and social-post agents. Barge-in, language switch, conversation QA.

Feedo

Ops staff describe a bug in Farsi. QA gets a classified ticket and a write-up they can paste into GitLab.

- Vision on the screenshot. Manager approve / reject. Next.js + FastAPI + Keycloak roles.
- Built for an internal ERP team — non-technical reporters, no ticket-form training.

LLM compression — Unsloth

Fine-tune and shrink models so they run on one GPU instead of a hosted API.

- LoRA / QLoRA on 4-bit bases. Domain multi-turn data. Measure the task, not a pretty loss curve.
- Distill a large teacher into a smaller student, then quantize and serve locally (Ollama / GGUF).

Persian STT eval

Whisper (FA) vs NeMo FastConformer on an RTX 4060 — pick the Farsi ear before meeting/voice prod.

How I ship

Full-stack 0 → MVP

AI title. Still write the product around the model.

- **Design** the loop (who types, what the model sees, when a human steps in).
- **Build** API + UI + auth + DB (FastAPI, Next.js/React, Keycloak, Postgres, RTL when needed).
- **Deploy** Docker / GitLab CI / Nginx → live URL. Then iterate with vibe-coded agent passes (Claude Code, Hermes, MCP).



THIS PAGE

EXPERIENCE

Danak · Ramzagah · freelance
WordPress

EDUCATION

Shiraz M.Sc. AI · Yazd B.Sc. CS

LANGUAGES

Persian	Native
English	Advanced

CONTACT

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

Experience + Education

Professional Experience

AI Developer Danak

Internal LLM features for a non-technical ops team — chat, document reading, automation.

- Wired models into daily workflows. Not a research stall — people used what shipped.
- Same habit as the products above: UI + API + a model, then hand it to staff.
- Scope: chat, document reading, and small automations for 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.

WordPress Developer Freelance · 2016 — 2024

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

- Client sites end-to-end. PHP, custom plugins, handoff — the muscle behind “I can MVP.”

Education

Shiraz University

2023 — present

M.Sc. Computer Engineering — Artificial Intelligence

Yazd University

2020 — 2023

B.Sc. Computer Science