



GERGŐ SÁRFFY

AI Engineer / Full-stack Software Developer

✉ hello@sarffy.dev 📞 +36 30 381 9422 📍 Tilburg, Netherlands 🌐 sarffy.dev 🔗 linkedin.com/in/sgeri

SUMMARY

Software engineer focused on **shipping full-stack products and applied AI systems**. Seven years of industrial-grade delivery across finance, energy, defense and contracting: **TypeScript full-stack web**, React Native apps, enterprise RAG and knowledge platforms, intelligent document processing and **fully self-hostable deployments** for organizations that cannot send their data to a public API. Comfortable owning features end to end from implementation through go-live.

Also invested in the practical side of AI inside engineering teams: writing internal AI usage policies, developer education and mentoring, and building agentic development systems and harnesses that turn **AI-assisted development** into a standard, reviewable part of the workflow. Based in **Tilburg, Netherlands**, available remote and across the EU.

SKILLS

Languages

TypeScript, JavaScript, Python, Go, SQL

Web & Application

React, Next.js, React Native, Expo, Tailwind CSS, NestJS, Node.js, Express, FastAPI, WebSockets, REST, queue workers, GraphQL, microservices

AI & LLM Engineering

RAG architectures, LLM inference infrastructure, vercel ai sdk, LangGraph, OpenAI Agents SDK, vLLM, model serving, embedding pipelines, hybrid search, reranking, agentic workflows, evaluation harnesses, guardrails & PII redaction, self-hosted LLMs

Data, Cloud & Infrastructure

PostgreSQL, MongoDB, Prisma, pgvector, Qdrant, Redis, Azure, AWS, Docker, CI/CD, GitHub Actions, Linux

Architecture & Leadership

architecture reviews, project scoping & estimation, API & integration design, hiring & leading subcontractors, code review, mentoring, CI/CD & engineering standards, technical proposals

WORK HISTORY

X Consulting

Software Engineering Lead

Budapest, HU / Remote

November 2025 - Current

Enterprise AI consultancy & software agency (x-consulting.co). Technical Lead.

- **Lead all technical projects** in the company, owning architecture, delivery and quality across the client portfolio from discovery call to production go-live.
- Built and shipped **three product lines** end to end: an enterprise AI knowledge base platform, an AI document validation platform and an AI pitch-deck generation platform, delivered to **key players in the energy and space/defense sectors** plus international consulting firms.
- **Hire, manage and lead subcontractor engineers**: scoping, task breakdown, code review and quality gates, on top of company-wide repository, CI/CD and environment standards.
- Work with client engineering teams on **API contracts, integrations and security/data residency reviews**; own client relationships from **technical pre-sales through to C-level reporting**.
- Convert business ideas into **implementation proposals** covering technical feasibility, effort and cost estimates and expected return, then lead the resulting SDLC.
- **Stack**: TypeScript (Next.js, NestJS, React, Expo), Python, Go, Azure, PostgreSQL, Vercel AI SDK, LangGraph, Qdrant, Docker.

1337 Partners

AI Engineer

Remote

January 2026 - August 2026

Contract - Agency & Applied Generative AI Platform Development

- Shipped smaller AI client use cases, including a **generative AI hotel customer service solution** and an **AI voice agent** built with ElevenLabs.
- Contributed to an **enterprise AI platform** with model deployment management, advanced user management and authentication, audit trails, admin dashboards and external data connectors.
- Supported **custom LLM inference work**: capacity estimation, GPU planning, model serving optimization and quantization across managed and self-hosted setups.
- Built **RAG pipelines**: chunking strategy, embedding model selection, hybrid retrieval with reranking, and an evaluation harness that caught retrieval regressions before release.

- Shipped **Python and Node.js workers and microservices** for asynchronous document, media and batch inference workloads, queue-backed and horizontally scalable.
- Delivered **agentic workflows** with tool calling and structured outputs, plus PII redaction and prompt-injection guardrails, keeping deployments portable and **self-hostable**.
- **Stack:** TypeScript (Next.js, NestJS), Python (FastAPI), Azure, vLLM, LiteLLM, OpenAI Agents SDK, Qdrant, Redis, Docker.

Peak Financial Services / PeakX
Software Architect
Full-time · Fintech platform (peakfs.io) / enterprise AI division (peakx.ai)

Budapest, HU
October 2024 - December 2025

- **Sole founding engineer of PeakX**, taking Peak's enterprise AI division from zero to production, including the RAG engine behind its AI proof-of-concept offering for banking and enterprise clients.
- Mapped and **reviewed technical feasibility of AI projects** for a large Hungarian bank, then led the implementation into production.
- Integrated LLM platforms with open-source and managed AI services (vector databases, embedding pipelines, multimodal generation) for enterprise and SME customers, laying the architecture of **AI solutions serving 10,000+ users**.
- Delivered small to mid scale end-to-end implementations, owning the lifecycle from planning, estimation and budgeting through to go-live.
- Helped establish a **development team of 7**, coordinated the SDLC across parallel projects and introduced coding standards, onboarding, development strategy and DevOps pipelines.
- **Stack:** TypeScript, Next.js, NestJS, React, RAG architectures, vector databases, Azure.

Peak Financial Services
Full Stack Developer
Full-time

Budapest, HU
June 2023 - October 2024

- Built proof-of-concept projects to assess client needs and implemented a shared design system in small dynamic teams.
- Experimented with **early AI technologies** and evaluated their viability in production, the work that grew into Peak's AI practice.
- Developed user-facing features in enterprise finance applications alongside external development teams and business representatives.

Rollout IT
Full Stack Software Engineer
Contract - Software Outsourcing

Remote
May 2023 - December 2025

- Delivered feature work and full-scale refactors across several inherited client codebases, improving maintainability and performance.
- Led the mobile-friendly rework of an **algorithmic trading platform**, making the core workflow usable on small screens.
- Built an **AI-powered used-car search engine**, applying machine learning to improve result relevance and engagement.
- **Stack:** TypeScript / JavaScript, React, Express, PostgreSQL, MongoDB, Supabase.

Wedding Manager
Full Stack Web & Mobile Developer
Part-time · Wedding planning marketplace

Budapest, HU
June 2021 - March 2023

- Early-stage **startup**: took part in product development, brainstorming and initial tech planning alongside day-to-day delivery.
- Built web and mobile features for a platform that **connects couples with venues and wedding service providers** and helps them plan the full wedding journey.
- **Stack:** React, React Native, NestJS, AWS.

Various clients
Freelance Web Developer

Remote
January 2019 - January 2022

- Delivered end-to-end web products from portfolio sites and WordPress e-commerce to custom **ERP and CRM applications** built with Next.js and MongoDB.
- Deployed and maintained everything personally on Vercel, PlanetScale and Railway, picking up Linux administration, networking and UI/UX along the way.
- **Stack:** React, Express, MongoDB, WordPress.

LANGUAGES

English C2 / Full Professional	Hungarian Native	German Limited Working	Dutch Beginner (<A1)
--	----------------------------	----------------------------------	--------------------------------