Arslan Jan Soomro

Senior Full-Stack AI Engineer

Islamabad, Pakistan · Open to new roles · Remote
arslanjs.dev@gmail.com · arslanjs.dev (case studies) · linkedin.com/in/arslanjs · github.com/Arslan-Soomro

Summary

Full-stack AI engineer who builds LLM products end to end (RAG, agents, evals) in TypeScript and Python on AWS. Tech lead on the **Pixis × Google co-development initiative**, now live as Google Creative Coverage. **Led a 6-engineer team** at Pixis, and raised a medical-coding RAG system's retrieval hit rate **from 28% to 75%**.

Skills

AI and LLM: RAG, AI agents, tool calling, LLM evals (LLM-as-judge, TruLens), prompt engineering

AI tooling: LangChain, LangGraph, Vercel AI SDK, Langfuse, OpenAI, Claude, Gemini, Pinecone, pgvector

Languages and frameworks: Python, TypeScript, SQL, React, Next.js, Node.js, FastAPI, tRPC

Data and cloud: PostgreSQL, MongoDB, Redis, AWS, AWS Bedrock, Docker, Kubernetes, Pulumi, CI/CD

Experience

Pixis AI · Engineering Lead

Remote · Apr 2026 – Oct 2026

Enterprise AI marketing platform. Led engineering for AdRoom, its AI ad-creative studio, plus ai-ad-gen and pixis.ai.

- Took the Pixis × Google co-development initiative **from prototype to production within two months** as tech lead; it's now live as Google Creative Coverage. Built the 5-step prototype in a week to align both teams.
- Led a 6-engineer team that shipped **4 major features in 4 months** (UGC video generation, ABCD video scoring, a video editor, credit-hold billing), reviewing and merging ~170 PRs across 3 production repos.
- Built ai-ad-gen, an agentic ad generator in ongoing production use by a Pixis enterprise client (25+ ads by handover): **about 8 in 10 images keeping product details intact** across eval batches, using a planner–producer pipeline with critic loops.
- Cut critical and high security advisories on the pixis.ai CMS **from 28 to 0** by upgrading Craft CMS and patching PHP, then moved it onto Pixis's own AWS account as code.

Stack: TypeScript, Python, FastAPI, React, Vercel AI SDK, Kubernetes, AWS, Pulumi

Case studies: arslanjs.dev/work/google-creative-coverage · /work/adroom · /work/ai-ad-gen

Claire IT AI · Lead Engineer

Remote · May 2025 – Oct 2026

AI medical-coding assistant SaaS with paid tiers and 146 accounts. Repo owner on a team of 3–4 contributors.

- Raised retrieval hit rate on pasted clinical charts **from 28% to 75%**, and cut failed searches from 10 to 0 on 20 coder-reviewed charts, by tracing and fixing five separate retrieval faults.
- Cut p90 answer latency **by 40% (83 s to 50 s)** and cost per answer by about 14% by removing duplicated context, fixing prompt caching and upgrading the AI SDK.
- Repaired **17 of 30** reviewer-flagged answer failures before users saw them by adding an answer-verification step.
- Built the eval harness that validated these changes (LLM-as-judge scoring, prompt A/B tests) and the RAG orchestrator over ~169k medical codes.

Stack: TypeScript, Next.js, Vercel AI SDK, Claude, AWS Bedrock Knowledge Bases, PostgreSQL

Case study: arslanjs.dev/work/claire-it-ai

Hammurabi · Founding & Lead Engineer — Remote · Apr 2024 – Mar 2026

AI legal-analysis SaaS for Israeli lawyers, in Hebrew and English.

- Took the platform **from first commit to production in about 3 months**, then grew it into a paid product, building solo for the first 10 months before leading the team.
- Shipped **~12 legal AI tools** to paying users, including argument analysis and contradiction detection, by building one config-driven framework they all run on.
- Picked the production retriever with evals, not guesswork: a sentence-window, linked-chunk design tested in TruLens against three baseline strategies on Hebrew and English legal text.

Stack: Next.js, tRPC, MongoDB, Redis, Pinecone, Python, FastAPI

Case study: arslanjs.dev/work/hammurabi

Digitenic (now Revonix) · Co-Founder & Lead Engineer — Islamabad, Pakistan · Jun 2024 – Oct 2026

AI engineering company building LLM products for clients. Led engineering on the products above.

- Tripled candidate screening **from 4 to 12 per 2 hours** by taking Proctor to production: candidates record answers asynchronously, replacing 30-minute live screening calls.
- Also shipped BudgetHub, a Swiss budgeting SaaS, and bootstrapped Pixis Visibility, an AI SEO platform.

Automychat & Othersify · Founder — Pakistan · Jun 2023 – Aug 2024

- Built a multi-tenant WhatsApp AI-chatbot platform for small businesses, and led a 5-person startup that shipped a Gmail read-receipt extension.

Dynalytix AI · Full Stack Developer — Jun 2023 – Aug 2023

- Top frontend contributor on Doclytix, an AI document-management SaaS: **292 of 593** frontend commits and 37 merged PRs in 8 weeks.

LeaseMagnets (Tour) · Full Stack Developer — Remote (Michigan, US) · Jun 2022 – Jul 2023

- **#3 contributor of 43** on Tour's React SaaS dashboard. Led its Next.js 13 rewrite and built billing, contracts and live co-browsing.

Education

Sukkur IBA University · BS Computer Science — Sukkur, Pakistan · Feb 2020 – Feb 2024