




# SAFZAN PIRANI

Agent Engineer

Hyderabad, India · Remote

✉ [safzanpirani@gmail.com](mailto:safzanpirani@gmail.com)  [safzan.dev](https://safzan.dev)  [github.com/safzanpirani](https://github.com/safzanpirani)  [linkedin.com/in/safzanpirani](https://linkedin.com/in/safzanpirani)

## Summary

---

Agent engineer shipping end-to-end — model, tool layer, infra, and human-in-loop surfaces. Comfortable owning the whole loop: Convex and Cloudflare Workers for state and scheduling, Postgres / Supabase for relational data, vLLM / SGLang / llama.cpp and ComfyUI for local inference, Telegram bots and operator consoles as human surfaces.

## Work Experience

---

### AltUni Labs @ InsideIIM

Sep 2025 – Present

Software Development Engineer

Remote

- Building agentic B2B hiring features end-to-end — model selection, tool layer, infra, and operator UI.
- Integrating LLM workflows into production with streaming responses and tool-use; owning the loop from prompt to deploy.
- Shipping low-latency real-time AI features behind real users; iterating on agent reliability under traffic.

### RadicalX

Dec 2023 – Feb 2024

Artificial Intelligence Engineer Intern

Remote

- Built ReX, an AI career-coaching platform, on OpenAI and Vertex AI with structured prompting and conversational flows.
- Reduced API spend by up to 60% via token-efficient prompts and response caching.
- Designed evaluation metrics and test harnesses to measure and improve LLM-generated response quality.

## Projects

---

### Spore | *Next.js, Convex, CF Workers, Flue*

Feb 2026

- Long-running task agent with a browser operator console — doom-loop guards, mid-prompt interrupters, follow-up steering, real-time token tracking, replay-based forking, and a subagent canvas via Flue's task tool.

### OpenPorts | *FastAPI, SQLite, React, Telegram*

Feb 2026

- Operator console + Telegram bot + recheck cron for publicly exposed ComfyUI / Ollama instances. Classic agent loop — scan, fingerprint, diff, notify — with URL-synced filters, saved searches, watch list, model catalog, and history + diff detection.

### Agentic Resume Editing | *TypeScript, Convex, Next.js*

May 2026

- Agentic LaTeX editor on Convex — live preview alongside an agent that edits the LaTeX tree on user instruction. Runs on self-hosted Convex behind a Cloudflare tunnel.

### Voicewin | *C#, .NET, Whisper, Deepgram*

Dec 2025

- Native Windows speech-to-text with global hotkeys — push / tap right-alt to record, transcribe via Whisper or Deepgram, paste. Ships signed releases.

### Flux Kontext Diff Merge | *ComfyUI Custom Node, Python ( 94+)*

2025

- ComfyUI node that preserves image quality on Flux Kontext edits — diffs the edit against the original and selectively merges only changed regions back, instead of degrading untouched pixels through a full re-encode.

## Technical Skills

---

**Agent Harnesses:** Flue, LangGraph, LangChain, Anthropic SDK, OpenAI SDK, Model Context Protocol (MCP)

**Languages:** TypeScript, Python, Go, C# / .NET, SQL, C++

**Backend & Runtime:** Next.js, FastAPI, Hono (Bun), Convex (incl. self-hosted), Cloudflare Workers, Cloudflare Tunnels

**Local Inference:** vLLM, SGLang, llama.cpp, ComfyUI, Whisper

**Operator Surfaces:** Telegram bots as control planes, CLI agents, browser operator consoles

**Frontend:** React, Tailwind, vanilla CSS, shadcn/ui

**Data & Infra:** Postgres / Supabase, SQLite, APScheduler, Docker, Git, VS Code / Zed

## Education

---

### St. Peter's Engineering College

2020 – 2024

Bachelor of Technology in Computer Science Engineering (AI/ML), GPA: 7.42

Hyderabad, India

### Kites Junior College

2018 – 2020

Intermediate Education

Hyderabad, India