# Shashwat Jain.

shashwatjain2k3@gmail.com
shashwatjain.tech
github.com/sahabji0P
linkedin.com/in/itsshashwatjain

AI ENGINEER · AIRA AI · BANGALORE / DELHI NCR

*Agents are my favourite thing to build, and most start as something I wanted for myself: an arena where my gems, my team of assistants, live; an eye on every token my coding agents burn; a piggy bank they can talk to. At AiRA, we're building a personal assistant you can call that remembers you, works in your apps and browser without seeing your passwords, and keeps going after you hang up.*

## EXPERIENCE

### AI Engineer · AiRA AI, Bangalore, India
Nov 2025 to Present

- Owned AiRA's AI infrastructure and led AI design decisions across the team; replaced Pydantic AI with a custom agent harness on raw SDKs across 7 providers, to control cache breakpoints, context loading and permission gates.
- Shipped the real-time voice runtime: cascaded (STT → LLM → TTS) and speech-to-speech pipelines with ~900 ms p50 time-to-first-audio and barge-in; ~5x smaller audio via 16-bit binary PCM over base64 floats.
- Designed AiRA's 4-tier memory system on Qdrant, Redis and MongoDB; a nightly pipeline detects open loops and drives follow-ups via push notifications, shipped with the iOS engineers.
- Analyzed per-provider caching and restructured prompts around a ~30k-token cross-user shared prefix with explicit cache breakpoints, sustaining a 98–99% cache-hit rate, cutting input-token cost ~90% and TTFT ~80%.
- Benchmarked and moved internal routing from generative LLM calls to a decision model, cutting per-decision latency from 600+ ms to under 150 ms, removing output-token cost and cutting input-token cost by over 50%.
- Set up prompt regression evals on passing and previously failing scenarios, shipping only when LLM-as-judge and human review agree; tracing via Langfuse and an in-house system.

### Full-Stack Developer · Web Landers, Remote
May 2024 to Oct 2025

- Led planning and work distribution for a 4-person team, delivering 5+ web apps (React, Next.js, Python, Supabase).

## SKILLS

| | |
|---|---|
| AI Engineering | LLM Orchestration, Multi-Agent, Tool Calling, RAG, Embeddings, Agent Memory, Prompt Caching |
| Model APIs | OpenAI, Anthropic, Gemini, xAI, DeepSeek, Groq, Cerebras, TypeSafe AI, OpenRouter |
| Agent Frameworks | LangChain, Pydantic AI |
| Voice AI | Cartesia, ElevenLabs, OpenAI Realtime, Grok Voice |
| Machine Learning | PyTorch, Hugging Face, Computer Vision (YOLO, Vision Transformers) |
| Languages | Python, TypeScript, SQL |
| Backend & Data | FastAPI, WebSockets, Redis, MongoDB, Qdrant |
| Cloud & DevOps | Azure, Gemini Enterprise Agent Platform (Vertex AI), Docker, GitHub Actions |

## PROJECTS

### Cortex · Arena for Gems
Jun 2026 to Present

Python · FastAPI · WebSockets · OpenAI · Anthropic · Cartesia · MCP · MongoDB · Redis · Rust (GPUI)

- Built a real-time voice cascade with barge-in that stores a cut-off reply as the words actually heard.
- One agent loop for calls and chat, with MCP tools opened per gem; folding in Vetaal and its vault.

### Gullak · AI Finance Assistant & MCP Server
May 2025 to Present

Python · MCP · Gemini 3.5 Flash (Vision) · Next.js · SQLite

- Implemented receipt-based bill splitting: Gemini 3.5 Flash vision turns bill photos into itemised expenses, cutting entry time from 5 minutes to 30 seconds.
- Built an MCP server exposing accounts, budgets and settlements as a narrow, validated tool set, so external agents can manage a user's finances.

### Suraksha AI · Real-Time Threat Detection
Mar 2024 to Jan 2025

Python · YOLOv11 · OpenCV · Flask · Roboflow · Firebase

- Curated a custom threat-detection dataset from Roboflow and Kaggle sources with augmentation for robustness, used to fine-tune YOLOv11 across 6 weapon classes.
- Developed the Flask backend for real-time inference on live video streams at 30 FPS, with Firebase alerts.

## EDUCATION

### B.Tech, Computer Science & Engineering (Artificial Intelligence)
2022 to 2026

Bennett University · CGPA 8.97

## PUBLICATION

**Brain Tumor Classification Using DeiT Vision Transformer** · 16th IEEE ICCCNT, IIT Indore — July 2025

## RECOGNITION

| | |
|---|---|
| Certifications | ServiceNow Certified Application Developer (2025); Microsoft Azure AI Fundamentals (2024); Malware Analysis and Introduction to Assembly Language, IBM (2024) |
| Achievements | Dean's List Award (9.8 SGPA, 2024); Hackaccino Hackathon Runner-Up (2024); Selected for Project Showcase 2024 among 100+ teams |