

Mokshit Jain

Bengaluru, India • +91-7000209021 • mokshitjain18@gmail.com • linkedin.com/in/mokshit-jain • github.com/mokbhai

EXPERIENCE

DrCode.ai / Airia Technologies Pvt. Ltd. • Software Development Engineer Apr 2025 – Present

- Designed a **stage-by-stage accuracy evaluation harness** for an AI compliance agent that reviews vendor submittals against engineering specs and produces **cited, structured findings**, so retrieval, extraction, and synthesis stages could each be scored on their own output instead of judging the pipeline only by its final answer.
- Built a **pgvector RAG pipeline with semantic caching** (LangChain/LangGraph, OpenAI embeddings) that cut **LLM call costs ~30%**, feeding a **deterministic extraction pipeline**: GPT-4o Vision pulls exact amount, party, and cheque-number fields off a photographed cheque, and a rules-based settlement step writes the result into the accounting system rather than letting the model generate free text. A human approve/disapprove step on every extraction feeds a party-name mapping table that improves recognition over time (**95%+ overall, 99% party resolution** over ~300 cheques in its first 10 days live).
- Designed and run a personal **multi-agent orchestrator** (*orca*, public, pipx-installable) that separates planning from execution: an orchestrating agent plans and writes small, scoped instructions, and parallel worker agents execute them, with durable run artifacts, per-task telemetry, and resumable sessions. Used it to modularize ~14k LOC a single agent session couldn't hold in context.
- Re-architected fragile cron pipelines into Temporal.io workflows with parallel execution: **0 production incidents** in the 2 months after migration. Maintained **>90% automated test coverage** on the backend (GitHub Actions CI/CD).
- Built an **AI-agent UI-testing toolkit** ground-up: Playwright exposed as an MCP server plus a Chrome extension that captures a tester's real clicks, with the agent **sandboxed to staging deployments** and never given the user's own credentials; cut manual QA effort **~80%**. Also shipped an API regression suite covering **120+ endpoints**, running on every PR.

Key systems (built or contributed to):

- YourMitra** (yourmitra.com; [case study]): *sole engineer, ground-up; live with paying customers*. Personalized AI story-book platform; multi-model generation (OpenAI, ElevenLabs, Replicate) parallelized with Temporal.io, cutting generation **20 min → 2 min**.
- FDC** ([case study]): WhatsApp cheque bot (Evolution API) built and integrated independently on a team project; compared **10+ OCR approaches** for extraction quality against cost before committing to one.
- AI-Assisted Interior Design & 3D Visualization Platform** (client; [case study]): multi-step AI pipeline (Gemini + OpenAI, 4 stages: dimensions → panels → assets → cove) feeding a headless Blender render and a 22-rule cost/materials engine; my contributions were color-differentiated materials, photorealistic rendering, resize and rule-engine fixes, and modularizing multi-thousand-line files.
- UPSCash** ([case study]): governed integration with QuickBooks Online (Intuit OAuth 2.0); AES-256-GCM token storage, HMAC-verified webhooks, RBAC, and audit logging over every reconciliation action.

Stack: Python, FastAPI, PostgreSQL/pgvector, Redis, Temporal.io, OpenAI, Gemini, LangChain/LangGraph, Docker, AWS, GCP

Innovation Studio LPU × Ministry of Defence • Freelance Developer Nov 2024 – Jan 2025

Jalandhar • Sole developer, requirements through deploy • recognized by the Indian Army (see Achievements)

- Built a secure vehicle-checkpoint management system tracking **10,000+ daily movements across 100+ checkpoints**; real-time license-plate recognition via a YOLOv5 ANPR model + OCR.
- Cut manual processing **48 hrs → 5 min (~85% efficiency gain)**; built real-time checkpoint-activity dashboards; met **100% of deadlines**.
- Multi-tier RBAC, MFA, and session-level security; tuned Prisma/PostgreSQL for high-throughput reads at peak.

Stack: Python, YOLOv5, Flask, Nest.js, Next.js, PostgreSQL, Prisma, Microservices

Jain Parichay • Software Developer Jun 2024 – Apr 2025

Marriage-biodata circulation platform •

- v1** (jainparichay.in, live): Express/EJS + Redis biodata-PDF circulation with Google-OAuth admin, Google Drive → S3 sync, token share-links, and a range-streaming PDF viewer (~38 endpoints); serving real monthly traffic.
- v2**: NestJS/Astro monorepo built around a **Temporal + LangGraph document-ingestion pipeline**: Azure Document Intelligence OCR over a submitted biodata PDF, a GPT-4o-mini stage that extracts structured fields, and a separate photo-extraction step, writing clean data to Postgres for the user to **review and correct before** it generates the final formatted profile: AI-extracted and human-reviewed content stay as distinct, auditable steps.

Stack: NestJS, Astro, tRPC, PostgreSQL, Redis, Temporal.io, LangGraph, OpenAI, Azure Document Intelligence, self-hosted S3

PERSONAL PROJECTS

orca • multi-agent CLI orchestrator

github.com/mokbhai/orca

Python • pipx • pluggable worker adapters

- A single-file Python CLI that fans work out to parallel headless AI worker agents (cmd, Cursor, Claude, Codex, Kimi, OpenCode), with durable run artifacts, telemetry, and resumable sessions; pluggable adapters so a new worker type can be added without touching the orchestrator core. I run it with Claude Code as the orchestrator: the expensive model plans and delegates, cheaper open-source models do the writing. [case study]

TECHNICAL SKILLS

Python [FastAPI], TypeScript [Node.js, NestJS], SQL

LLM APIs [OpenAI incl. Vision, Gemini, Azure Document Intelligence, LiteLLM], prompt engineering

RAG [LangChain, LangGraph], vector search [pgvector, ChromaDB, Pinecone], semantic caching, multi-agent orchestration

Deterministic extraction pipelines, OCR, evaluation harnesses, per-stage accuracy measurement, production feedback loops

RBAC, JWT/OAuth2, AES-256-GCM token storage, HMAC-verified webhooks, audit logging

PostgreSQL, Redis, MongoDB, Temporal.io, Docker, AWS [S3, EC2], GCP [Compute Engine], CI/CD, Sentry

Pytest, Playwright, Git, Linux, Agile/Scrum; DSA [350+ solved], Low-Level Design, DBMS, OS

EDUCATION

B.Tech, Computer Science & Engineering • Lovely Professional University

GPA: 8.1/10 •

CERTIFICATIONS & ACHIEVEMENTS

- Generative AI: Introduction and Applications, Coursera • Prompt Engineering, Coursera • Complete Machine Learning & Data Science Program, GeeksforGeeks.
- Certificate of Appreciation, Indian Army (387 (I) Sup Pl ASC) for a vehicle-checkpoint automation project: real-time YOLOv5 ANPR and OCR over **10,000+ daily movements**, manual processing **48 hrs → 5 min** • Certificate of Recognition, DRCODE, for ownership and production incident resolution.