

Harshit Pandey

✉ contact@harshitpandey.com ☎ +918287253559 🌐 LinkedIn 🏠 Github 🌐 Website

📁 EXPERIENCE

Software Development Engineer II, InteligenAI

10/2025 – Present | Gurugram, India

Financial Document Intelligence

- Built a document AI pipeline for mortgage files: process-pool rendering, batched OCR, pgvector (HNSW) page classification and parallel Gemini extraction.
- Cut LLM token spend by 45% and p95 extraction latency from 3.2s to 680ms with continuation batching. Later batches receive a compact summary of captured fields and run in parallel.
- Shipped hybrid BM25 + vector page classification and a deterministic, LLM-free standard form extraction engine with a precision-first fallback to Gemini.
- Managed GCP infrastructure in modular Terraform across dev, stage and prod (Cloud Run, Postgres, Redis, Secret Manager, monitoring), sustaining 99.95% uptime.

Industrial Telemetry Pipeline

- Built an end-to-end pipeline from 20ms-aligned PLC acquisition through FastAPI and Redis Streams consumer groups to TimescaleDB, handling ~10,000 tag values/sec with bounded buffering and per-destination retries.
- Indexed telemetry in Elasticsearch for full-text and analytical search. Built Grafana, Loki and OpenTelemetry dashboards with automated alerting on pipeline health.

Platform & Team

- Architected and published a TypeScript/Node.js SDK for real-time messaging integrations, standardizing retries, backoff and signature verification.
- Led a team of engineers, establishing code review standards and CI/CD practices that improved deployment frequency by 30%.
- Designed and built a reusable agent platform (LangGraph, MCP, LLM gateway) with versioned skills, per-user access control and human approval gates; the industrial reliability agent runs on it.

Software Development Engineer I, InteligenAI

08/2024 – 10/2025 | Gurugram, India

- Designed a multi-agent reliability assistant (LangGraph) that answers root-cause questions from a knowledge base of equipment documents and RCAs, live ML alerts and sensor history, implemented by a team of engineers.
- Developed an AI-powered visa processing engine in Python and Node.js to automate regulatory workflows.
- Built an ARIMA sales forecasting platform and a patient portal (IPD/OPD) covering scheduling, medical tracking and admissions.

SDE Intern, InteligenAI

01/2024 – 07/2024 | Gurugram, India

- Optimized PostgreSQL queries through EXPLAIN ANALYZE profiling, composite indexes and JSONB refactoring, cutting p99 API latency by 42%.
- Built a schema-driven form configuration engine with validation and webhooks, cutting client onboarding from 2 weeks to 2 days.
- Engineered payment gateway workflows for recurring billing, webhook verification and reconciliation.

🧠 SKILLS

Languages & Runtimes — Node.js (TypeScript, Fastify, Express), Python (FastAPI, Django), React, Next.js, SQL, C++

Backend & Distributed Systems — Multithreading/Multiprocessing, AsyncIO, Redis (Streams, Consumer Groups, Redlock), RabbitMQ, Celery, REST, RPC, WebSockets

Databases & Storage — PostgreSQL (Partitioning, GIN/B-Tree Indexing), TimescaleDB, Elasticsearch, MongoDB, pgvector

GenAI & Agentic Systems — LLM Orchestration, LangGraph, Model Context Protocol, Hybrid Retrieval (BM25 + Dense Vectors), Structured Outputs (Pydantic/Instructor), LLM Gateways (Bifrost)

Infrastructure & Observability — GCP (Cloud Run), AWS (ECS, Lambda, RDS, S3), Azure, Terraform, Docker, Kubernetes, GitHub Actions, Nginx, Grafana, Loki, OpenTelemetry

📁 PROJECTS

Reusable Agent Platform 🗑️, Python, FastAPI, LangGraph, MCP, Bifrost, PostgreSQL/pgvector, Keycloak, Langfuse, Docker

- Self-hosted platform running AI agents for multiple use cases behind one LLM gateway; the industrial reliability agent runs on it.
- 32-tool MCP server with per-tool authorisation and Postgres row-level security. Versioned, access-controlled skills loaded on demand.
- Human approval gates on risky actions, budgeted sandbox for scheduled agent code, agent output rendered on an isolated origin, and full tracing in Langfuse.
- 16-service Docker deployment (Keycloak sign-in, object storage, nightly backups, TLS proxy).

🎓 EDUCATION

Harcourt Butler Technical University, B.Tech - Electronics Engineering (CGPA- 7.92/10)

2020 – 2024