

Aayush Kumar Dubey

Bachelor of Technology (B.Tech)
Instrumentation and Control Engineering
NSUT, New Delhi

+91-8700770566

aayushdubey.work@gmail.com

GitHub Profile

LinkedIn Profile

EDUCATION

• Netaji Subhas University of Technology, New Delhi (Formerly NSIT)

2023 - Present

Bachelor of Technology in Instrumentation and Control Engineering

EXPERIENCE

• Software Engineering Intern

Jun 2026 - Present

Flywheel

- Built **Keycloak-backed JWT authentication and RBAC** from scratch, gating all mutating endpoints behind scoped permissions and adding tenant isolation across the API; verified by 15 auth and tenancy tests.
- Refactored the vector store behind a single **VectorStore protocol**, removing all **11 direct Marqo call sites** and **152** lines of duplicated logic, with test-enforced layering.
- Automated Langfuse evaluation pipeline for **679 real-world farmer queries**, achieving **100% validation success**, eliminating **122 duplicate test cases**, and improving production QA workflows.
- Migrated OCR to **self-hosted Chandra/Gemma on vLLM (H100)**, built an automated benchmarking harness, scoring **5 engines** by ANLS/CER/WER plus a vision-LLM judge.
- Diagnosed a scoring-fairness bug that penalized markdown formatting and understated the self-hosted engine by **2.4× (0.33 to 0.79 mean ANLS)**, then set the production engine choice on an accuracy-versus-latency tradeoff of **3s against 150s per page**.
- Designed **document lifecycle APIs** with fail-closed ordering between the vector index and relational store so deleted content is never left retrievable; covered by 13 unit tests and a Dockerized smoke stack.

PROJECTS

• KAI — Knowledge Augmented Intelligence | [Github](#) | [Live](#)

Apr 2026

Next.js, TypeScript, Gemini, Qdrant, Postgres, Docker

- Built a full-stack RAG workspace as a single Next.js App Router app where users upload PDFs and chat against their library, deployed on Vercel with Neon Postgres and Qdrant Cloud.
- Implemented hybrid retrieval — dense vector plus keyword search fused with Reciprocal Rank Fusion and a Gemini reranker — so every streamed answer carries page-level citations and inline PDF preview.
- Added multi-document selection, cookie-based guest workspaces, and health checks that degrade cleanly when external services are unset.

• FusionX-3: AI-Generated Image Detection

Sep 2025

PyTorch, Xception, CNN, Digital Image Forensics

- Developed a **triple-stream CNN** combining RGB, DCT, and PRNU forensic features, trained on **120,000 images** and reaching 95% accuracy and a 0.95 F1-score.
- Implemented a patch-wise Xception architecture with 1×1 convolutional classifiers for localized artifact detection, reducing overfitting and minimizing validation loss to 0.12.

TECHNICAL SKILLS AND INTERESTS

Languages: C++, Python, SQL

Backend: FastAPI, REST API Design, Temporal, SQLAlchemy, Alembic, Redis, JWT, Keycloak, RBAC

Databases: PostgreSQL, MySQL, SQLite, MinIO/S3, Qdrant, Marqo

ML & AI: RAG, Hybrid Search, LLM Evaluation, LangChain, PyTorch, OCR, NLP

Infrastructure: Docker, Docker Compose, Git, Vercel, Render, vLLM, NVIDIA H100

CS Fundamentals: Data Structures & Algorithms, OOPs, DBMS

POSITION OF RESPONSIBILITY

• Logistics Coordinator — MOKSHA 2025 - The Annual Cultural Fest of NSUT

March 2025

- Coordinated logistics for 300+ volunteers and delivered all sponsor activations for the flagship festival.