Dev Jhawar

Bhubaneswar, India | +91-9733177667 | dev.jhawar.cs@gmail.com
linkedin.com/in/dev | github.com/dev

Education

Kalinga Institute of Industrial Technology — Bhubaneswar
Bachelor of Technology in Computer Science – CGPA 9.65 — *Aug. 2023 – Present*

Experience

IOT LAB, Technical Community Team — Nov 2024 – Dec 2025
KIIT University — *Bhubaneswar*

- Spearheaded **sponsorship outreach and partnerships** for lab initiatives, coordinating with **3+ external organizations** to secure resources and drive visibility for AI research projects.
- Organized and led **2 technical events** and cross-functional brainstorming sessions, managing end-to-end logistics and facilitating collaboration across **15+ team members** in a fast-paced lab environment.
- Supported **LangChain-based AI workflows** across lab projects by coordinating documentation, research pipelines, and internal knowledge sharing between the ML and non-technical teams.

Projects

PolicyLens – RAG Policy Analyzer *Live Demo* | *Python, FastAPI, LangChain, Supabase, Docker* — Jan 2026

- Built a **massively scalable** RAG pipeline with distributed vector storage (Supabase pgvector) and PyMuPDF ingestion, turning big data into action by parsing documents in **under 0.2s** for semantic search.
- Architected **low latency infrastructure solutions** by engineering **6 FastAPI endpoints** with async streaming and a CrossEncoder reranker, narrowing candidates from **8 to 5** for precise, real-time customer responses.
- Containerized via **Docker** with a **5-table** relational schema handling vector storage and chat history, ensuring robust, fault-tolerant deployment for continuous data streaming.
- Designed the end-to-end **query pipeline** – embedding user queries, retrieving semantically similar chunks, and streaming LLM responses through **LangChain** to seamlessly turn data into actionable answers.

LLM Semantic Cache *Live Demo* | *Python, FastAPI, Redis, PostgreSQL, Docker* — Jun 2026

- Built a caching layer for LLM APIs that checks if a similar question was already answered – reducing repeated LLM calls by **˜55%** by matching new queries against stored ones at a **0.85 similarity threshold** using embeddings.
- Used **Redis** to store cached answers with a **24-hour auto-expiry** and to limit each user to **10 requests per minute**, cutting average response time from **3.5s to 1.5s** on repeated queries.
- Built **3 API endpoints** – ask, analytics, and cache – with a **PostgreSQL** log tracking hit rate, response time, and cost saved per request, deployed live on Render with Supabase and Upstash Redis.

NetLink – P2P Terminal Messenger | *C++, TCP Sockets, POSIX Threads* — Jun 2026

- Developed a **real-time terminal messaging** application in C++ using **TCP sockets** for peer-to-peer network communication.
- Integrated **POSIX multithreading** to manage **concurrent data streams**, eliminating I/O blocking and enabling seamless two-way chat.
- Engineered **robust error handling** to detect network drops, **gracefully manage disconnects**, and ensure clean system exits.

Technical Skills

Languages: Python, Java, C/C++, SQL, JavaScript, HTML/CSS
Frameworks & Libraries: Scikit-learn, TensorFlow, PyTorch, FastAPI, LangChain, Pandas, NumPy, OpenCV
Dev Tools: Git, Docker, Redis, AWS, Postman, Google Colab, VS Code
Key Concepts: Data Structures & Algorithms, Object-Oriented Design, Low Latency Systems, Distributed Systems, Relational Databases, OS, Computer Networks, Agile Development

Achievements

- **Competitive Programming:** Innovated and problem-solved **900+ DSA problems** across LeetCode, CodeChef, and Codeforces – demonstrating mastery in **algorithm design** and **complexity analysis** to architect smart solutions.
- **Certifications:** Earned professional credentials in Data Analysis with Python (IBM) and Neural Networks and Deep Learning (DeepLearning.AI).