Harsh Kumar

☎ 9341045191 | ✉ harsh.k22@iiits.in | in LinkedIn | harshrajput4343 | Portfolio

Education

Indian Institute of Information Technology, Sri City Chittoor, India
Bachelor of Technology - Electronics & Communication Engineering; CGPA: 7.5 *Nov 2022 - June 2026*

Experience

CodeClause
Data Science Intern

- Configured AWS S3 buckets (raw/labeled/processed data), SageMaker Ground Truth labeling, 30+ annotated images.
- Deployed Lambda functions for preprocessing, configured IAM roles, orchestrated workflows via Step Functions.
- Built production-ready ML pipeline integrating S3, Lambda, Step Functions for end-to-end automation.

Research Assistant
Python and AI Backend Development

- Deployed FastAPI production APIs for IoT anomaly detection with CNN/attention-based models, 10K+ daily requests.
- Created scalable data preprocessing pipelines with NumPy and Pandas for high-volume sensor and network data.
- Integrated CNN, attention models, and XGBoost into backend workflows for real-time prediction and monitoring.
- Developed MLflow experiment tracking and model versioning pipelines for reproducible and reliable deployments.

Projects

FlowLoG (GitHub)
Full-Stack Kanban Project Management & Task Tracking Platform

- Built Kanban board with **Next.js 14**, **TypeScript**; integrated drag-and-drop and real-time task management.
- Architected RESTful APIs with **Express.js** and **Prisma ORM** for relational data across boards and cards.
- Designed card modals with **checklists**, labels, and member assignment for streamlined collaboration.
- **Technologies**: Next.js, TypeScript, Express.js, Prisma ORM, PostgreSQL, CSS Modules, @hello-pangea/dnd

MultiGPT (GitHub)
Multimodal LLM Router with Full-Stack Chat Interface

- Architected intelligent chat platform routing queries across **6 AI models** (Gemma, Nemotron, GLM) with keyword + AI-powered auto-routing and **real-time streaming** responses.
- Built full-stack app with JWT auth, RLS policies & **15+ CRUD ops**, reducing query response time by **40%**.
- Shipped **5+ production features** — tag search, sharing & markdown rendering with syntax-highlighted code blocks.
- **Technologies**: React 19, TypeScript, Vite, Supabase, PostgreSQL, OpenRouter API, Zustand, Tailwind CSS, Vercel

AI-Powered Medical Chatbot (GitHub)
LLM-Driven Healthcare Assistant with RAG and API Deployment

- Built AI-powered medical assistant using **GPT-4**, **Gemini**, **RAG**, **Pinecone** enabling context-aware diagnosis.
- Engineered scalable **LangChain** pipelines for document ingestion and semantic retrieval achieving 95% accuracy.
- Deployed **Flask** APIs with caching, logging, and error handling for scalable inference serving 1000+ requests.
- **Technologies**: Gemini, GPT-4, LangChain, Pinecone, Hugging Face, Flask, AWS, GitHub Actions, Embeddings

Technical Skills

Languages: Python, JavaScript, TypeScript, C/C++
Backend & Databases: FastAPI, REST APIs, Django, Node.js, REST APIs, PostgreSQL, MongoDB, Redis
Frontend: React, Next.js, TypeScript, HTML, Tailwind CSS, Real-time Dashboards, Data Visualization
Cloud & DevOps: AWS (Lambda, S3, SQS, CloudFront), Docker, CI/CD, GitHub Actions, Grafana
AI / LLM: APIs, ML, DL, MPC, LangChain, Vector DBs, RAG, GENAI, Pinecone, Prompt Engineering

Achievements & Certifications

- Oracle Cloud Infrastructure 2025 Certified Generative AI Professional *Link*
- Oracle Cloud Infrastructure 2025 Certified AI Foundations Associatel *Link*
- Solved 250+ problems on LeetCode/GfG with focus on algorithms, data structures, and problem-solving.

Leadership & Responsibilities

IIIT Sri City Cricket Team September 2024 – November 2025

- Led team strategy, match planning, and on-field decision-making during intercollegiate tournaments under high-pressure conditions.

Events Coordinator at Abhisarga March 2024 – December 2024

- Coordinated logistics and execution for 15+ technical events with 1,500+ participants, ensuring on-time delivery.