Ayush Shukla

Greater Noida, India | +91 9548902855 | ayushshukla6917@gmail.com | LinkedIn | GitHub | Website

B.Tech CSE '27 | Building Agentic AI, RAG & Real-Time CV Systems with FastAPI & LangGraph

EDUCATION

Galgotias College of Engineering and Technology (Affiliated to AKTU) Uttar Pradesh, India

Bachelor of Technology, Computer Science & Engineering *Expected May 2027 | CGPA: 7.3*

Relevant Coursework: AI, Machine Learning, Data Structures & Algorithms, OS, DBMS, Probability & Statistics, Linear Algebra.

WORK EXPERIENCE

HexSoftwares Remote, India

AI Engineer Intern *Oct 2025 – Dec 2025*

- Architected a production Retrieval-Augmented Generation (RAG) pipeline using Pinecone and Hugging Face, reducing query latency by 42% (120ms to 70ms) through batch ingestion, optimizing vector indexes, and asynchronous retrieval across 1000+ documents under concurrent workloads.
- Engineered a fault-isolated asynchronous FastAPI backend with separated embedding, retrieval, and inference microservices, adding timeout handling and fallback logic to prevent inference failures from impacting client-facing endpoints.
- Containerized microservices via Docker, streamlining deployment and reducing environment-related failures during testing.

TECHNICAL SKILLS

- **AI** / **ML**: PyTorch, LangChain, LangGraph, Multi-Agent Systems, Transformers, Hugging Face, Prompt Engineering, scikit-learn, NumPy, Pandas
- **Computer Vision**: OpenCV, YOLO, MediaPipe, MobileCLIP, Real-Time Video Processing, Object Detection, Pose Estimation
- **LLMs & Model Training**: QLoRA, LoRA, DPO Dataset Construction, Hugging Face Trainer, MLflow, Fine-Tuning, Whisper (ASR)
- **RAG & Databases**: Vector Databases (ChromaDB, Pinecone), Retrieval-Augmented Generation, SQL, PostgreSQL
- **Backend & Infrastructure**: FastAPI, REST APIs, Async Programming, Docker, Rate Limiting, Git/GitHub, n8n
- **Languages**: Python (Advanced), C++, C#

PROJECTS

DistribuFlow — AI-Powered B2B Order Management WhatsApp Agent Code | *July 2026 – August 2026*

Python, FastAPI, LangGraph, PostgreSQL, n8n, Llama 3.3, Whisper

- Engineered a stateful WhatsApp agent using LangGraph and an LLM to autonomously process multi-modal (text/voice) B2B orders with strict inventory and credit validation.
- Architected a deterministic product resolution engine via IDF-weighted fuzzy matching against a PostgreSQL catalog, effectively reducing LLM hallucinations.
- Built a decoupled FastAPI backend and n8n event-driven workflow to automate manager approvals, real-time admin alerts, and PDF invoice delivery.

AURA — Vision-Language AI Agent with Voice & Gesture Control Code | *Dec 2025 – Jan 2026*

Python, OpenCV, MediaPipe, Llama 3, Whisper, LangGraph

- Built a real-time multimodal AI agent integrating computer vision (OpenCV + MediaPipe) and voice (Whisper + LLM APIs), achieving 60 FPS via threaded pipelines and non-blocking inference.
- Implemented gesture-driven OS control (cursor, typing, app navigation) with temporal smoothing and noise filtering for stable interaction under low-light and rapid-motion conditions.
- Designed an asynchronous voice pipeline with LangGraph tool-calling architecture, enabling concurrent speech processing and system actions without blocking the vision pipeline on standard CPU-only hardware (<4GB RAM).

AgriFix AI — Multimodal Agentic Agri-Repair Assistant Code | *Oct 2025 – Jan 2026*

Python, FastAPI, Flutter, ChromaDB, Whisper, MobileCLIP, RAG

- Developed a multimodal AI agent delivering live AR repair guidance from farmer voice and image inputs, optimized for robust inference under low-bandwidth field conditions.
- Reduced end-to-end latency by 50%+ through RAG optimization and multi-hop inference under low-bandwidth field conditions.
- Built a production-ready FastAPI backend with rate limiting, input validation, and prompt-injection filtering to defend against malformed or adversarial inputs; containerized and deployed via Docker for portability.

ACHIEVEMENTS & LEADERSHIP

- **Team Lead** — Smart India Hackathon (SIH) 2025; led team through internal selection to become National Representative.
- **1st Place Winner** — HackOverflow v9.0 (2025) among 600+ national finalists for AI-powered AR system.
- **Finalist** — India Innovates (Asia's largest hackathon, 20,000+ competing teams).
- **Top 5** — Axis HackSprint (National).
- **Finalist** — CosmoHack1, DataVerse, BITS Pilani TechSeva.