

UZEB KHAN

New Delhi, India | Email | LinkedIn | GitHub | X | Portfolio

Summary

Junior year CS student who'd rather build than just study and picks real problems to solve, figuring out the right technologies and system design along the way, and fixing what breaks. Learns by shipping, not by reading docs cover to cover. **Targeting:** AI Engineer / LLM Engineer internships.

Education

KCC Institute of Technology and Management

Bachelor of Technology in Computer Science (Data Science) | AKTU

Greater Noida, UP, India

Aug. 2024 – 2028

Technical Projects

AXIOM — Multi-Agent STEM Solver | *LangGraph, FastAPI, SymPy, Qdrant, FAISS, Langfuse* **June 2026**

- Architected a LangGraph orchestrator that fans out specialist agents **in parallel** via the **Send API**, cutting end-to-end latency versus sequential execution.
- Built a hybrid **Qdrant + FAISS** retrieval layer with response caching and session-aware context for grounded, multi-turn document QA.
- Designed a **two-stage verification pipeline** (symbolic validation + LLM critique) that caught and fixed live bugs, including a critic bypass on high hop counts.
- Shipped a **FastAPI** backend and a **Streamlit** UI with live progress streaming, LaTeX rendering, and end-to-end **Langfuse** tracing.

SIFT — Autonomous Research Agent | *Python, Groq, OpenRouter, Tavily, Asyncio, FastAPI* **July 2026**

- Built a **ReAct-pattern** agent with **dual-LLM routing** : Groq for tool-calling, OpenRouter for summarization to balance speed and quality per task.
- Engineered a **three-tier fallback chain** (native tool calling → structured JSON → deterministic heuristics) so the agent stays functional through provider outages.
- Built an async crawling pipeline over Tavily search with **bounded concurrency** and citation tracking for traceable multi-source summaries.

Recall — Memory-First RAG System | *Python, LangGraph, Chroma, Sentence Transformers, Langfuse* **July 2026**

- Built a memory-first retrieval pipeline that checks episodic memory before full document search, cutting repeated-query response time by up to **180×**.
- Designed **confidence and freshness scoring** to auto-invalidate stale memory entries when source documents change.
- Added a **reconciliation node** to resolve conflicting memory states across sessions; audited and fixed bugs in routing logic and session locking.

Technical Skills

Languages: Python, C, C++, JavaScript, SQL

Agentic AI: LangGraph, LangChain, CrewAI, Groq, OpenRouter, Qdrant, Chroma, FAISS, Langfuse, RAG, MCP, Context Engineering

ML / Fine-Tuning: PyTorch, Unsloth, TRL, Hugging Face, scikit-learn, NumPy, Pandas, SymPy

Backend & Infra: FastAPI, Asyncio, Streamlit, PostgreSQL, MongoDB, Redis, Docker, Git, Linux, AWS (EC2, S3)