

SUMAMA BIN TAHIR

Islamabad, Pakistan | sumamatahir.ai@gmail.com | 0336-8491234 | [linkedin.com/in/chaudharysumama](https://www.linkedin.com/in/chaudharysumama) | github.com/ScriptSorcerer23

SUMMARY

AI Engineer specializing in LLMs, RAG pipelines, and agentic AI systems. Experienced building document processing workflows, embedding-based vector search, and SQL data pipelines, with additional skill in production-grade mobile app development and a track record of shipping AI solutions from prototype to production.

EXPERIENCE

- AI Intern (Team Lead) — National Centre of Artificial Intelligence**
Medical Imaging & Diagnostics Lab, Islamabad
Sep–Oct 2025
- Led a cross-functional team of 9 to build SearchDr, an agentic doctor-discovery system, cutting specialist search time by 90%.
 - Architected an agentic data pipeline with LangGraph, Playwright, Crawl4AI, and Llama 3.1 8B to extract, clean, and embed real-time specialist data for LLM-powered retrieval.
 - Designed the end-to-end agent workflow (ingestion, processing, response generation) for instant, personalized doctor recommendations from symptoms, reports, and location.
- Junior AI Engineer — Investince**
Remote
Jun–Sep 2024
- Integrated Mistral Large 2 into a real-estate geospatial intelligence platform (US/Canada), enabling natural language queries over OSM and Zillow data via LangChain.
 - Built backend AI services connecting LLM inference with real-time geospatial pipelines using Railway and Postman; scaled 3 open-source-model data extraction pipelines to 1M+ records/month.
 - Engineered LLM-driven pipeline automations that boosted geospatial data processing efficiency by 40%.

PROJECTS

- SolMate — Final Year Project** — Full-stack vitamin D and sun-exposure wellness app (React Native, Supabase, Flask) built around a LightGBM regression model trained on 37K NHANES records and a 94%-accuracy MobileNetV3 skin-type classifier. Includes a gamified community platform (challenges, leaderboards, streaks) and Solara, a LangChain/Gemini/Llama chatbot with Azure Neural TTS and Groq Whisper voice support.
May 2025–May 2026
- DocMind — Agentic RAG System** — Document intelligence platform (React, FastAPI, Supabase pgvector) where a CrewAI agent decides when and how to retrieve, exposing semantic search as a callable LangChain tool instead of retrieving on every message. Returns grounded, cited answers from uploaded PDFs/text files; deployed via Railway and Netlify.
May 2026
- Multi-Agent Defense OSINT Researcher** — Multi-agent research platform (LangGraph, MCP, Groq/LLaMA) for defense and aerospace intelligence: Search, Extraction, Synthesis, and Critique agents query verified sources (Janes, Aviation Week, ISPR), fact-check findings, and generate cited reports streamed live to a React UI over FastAPI WebSockets.
Jun 2026
- UniWeek** — University event management platform (React Native, Supabase, TypeScript) with AI-powered event recommendations and sentiment analysis for student engagement and society analytics.
Nov 2025
- Real-Time News Summarizer (RAFT)** — Retrieval-Augmented Fine-Tuning pipeline fine-tuning DistilBART on news data, with SentenceTransformers + ChromaDB semantic retrieval, served via FastAPI and Streamlit.
Jul 2025
- Taxi Teller NYC — MLOps Pipeline** — End-to-end ML pipeline on AWS (S3, EC2, EMR) processing 2M+ records with PySpark and MLflow; payment-prediction and fare-forecasting models deployed via Streamlit with CI/CD through GitHub Actions and Docker.
Jun–Jul 2025

EDUCATION

- BS Data Science — COMSATS University Islamabad**
Islamabad • CGPA: 3.30/4.00
Sep 2022–Jun 2026

CERTIFICATIONS & AWARDS

- LangChain & Vector Databases in Production (ActiveLoop) • AWS Cloud Practitioner Essentials (AWS)
- 3rd Runner-Up, Datathon Competition, COMSICON '25 • Shortlisted 2nd of 23 projects, FYP Expo — COMSATS University Islamabad

SKILLS

- AI/ML & Agentic Systems:** LangChain, LangGraph, CrewAI, LlamaIndex, MCP (Model Context Protocol), Agentic AI, RAG, Multi-Agent Systems, Prompt Engineering, Fine-tuning, LLMs (GPT, Claude, LLaMA, Mistral, Gemini, Groq)
- Vector Databases:** Pinecone, FAISS, Weaviate, ChromaDB, Supabase pgvector
- ML & Data Science:** Scikit-learn, PyTorch, OpenCV, MLflow, LightGBM, MobileNetV3
- Cloud & DevOps:** AWS (S3, EC2, EMR), Docker, CI/CD, Git, GitHub Actions, Railway, Netlify
- Programming, APIs & Frontend:** Python, JavaScript/TypeScript, Node.js, FastAPI, Flask, WebSockets, REST APIs, React Native, React, Expo
- Data Engineering:** Pandas, NumPy, SQL, PySpark, PostgreSQL, MySQL, Supabase, MongoDB