

Syed Muhammad Sarim Ahsan

+92 323 703 4858 | sarim.ahsan101@gmail.com | [Github](#) | [LinkedIn](#) | [HuggingFace](#) | [Portfolio](#)

SUMMARY

Computer Engineering student (GPA 3.5, SSUET) and Software Engineer building production full-stack platforms and agentic AI systems end-to-end — from multi-agent LangGraph pipelines and RAG applications to deployed, CI/CD-driven web platforms serving hundreds of live users. Comfortable owning a project from architecture through deployment: FastAPI/Flask and React/Node backends, Docker-based infra, and LLM-powered agents integrated into real products rather than notebooks.

EDUCATION

Sir Syed University of Engineering & Technology

B.S. in Computer Engineering · GPA: 3.5 / 4.00 · 6th Semester

Karachi

2023 – Present (Expected 2027)

Beaconhouse School System

Computer Science · A Grade

Karachi

2009 – 2021

TECHNICAL SKILLS

AI / ML: PyTorch, LangChain, LangGraph, HuggingFace, RAG, QLoRA, LoRA, scikit-learn, bitsandbytes

GPU / Systems: CUDA C++, custom kernels (tiled matmul, fused softmax), CUDA profiling

Backend: FastAPI, Flask, Node.js, Express, REST APIs, MySQL, MongoDB, Qdrant, ChromaDB, FAISS, Neo4j

Frontend: React, Vite, Tailwind CSS, HTML5, CSS3

Languages: Python, JavaScript (ES6+), C++, Java

DevOps: Docker, Docker Compose, Nginx, GitHub Actions, VPS / Linux

EXPERIENCE

Software Engineer

2025 – Present

Global Integration System (GIS, UK)

- Built and maintain [portalersm.com](#), a production student recruitment platform with secure RBAC serving 600+ active users across the UK, including full-stack feature development and zero-downtime CI/CD via GitHub Actions.

Machine Learning Engineer Intern

Jun 2026 – Present

FlyRank

- Working on applied AI and ML-based solutions as part of the FlyRank AI Internship Program in a fast-paced engineering environment.

Machine Learning Engineer Intern

May – 2026

Evello Pathways

- Built deployment-ready ML pipelines (regression, classification, clustering, recommendation, forecasting, computer vision) with Scikit-learn, XGBoost, PyTorch, YOLOv8, FastAPI, Docker, and MLflow.

IT Intern

2025

IFFCO Pakistan

- Audited 50+ enterprise hardware assets, documented network infrastructure, and assisted with endpoint provisioning, network troubleshooting, and IP configuration.

PROJECTS

Agentic AI & Applied AI Products

Chronicle — AI Repo Intelligence Agent | [Github](#) | [Live](#)

2026

- Agentic chatbot that ingests a repository's full commit, PR, and issue history into Neo4j to enable natural-language queries and surface risk metrics like "bus factor"; stack: FastAPI, Neo4j, Groq (LLaMA 3.3), React/Vite — deployed via Docker Compose with multi-repo switching and a commit diff inspector.

ArgoFarm — AI Crop Diagnostics Platform | [Github](#)

2026

- Full-stack AgriTech platform: Gemini 2.5 Flash Vision for real-time crop disease diagnosis, Groq LLaMA advisory chat, Groq Whisper voice input, live outbreak radar, and a bilingual (Urdu/English) marketplace — built with Flask, MySQL, TensorFlow, Vite, and Docker, deployed on Netlify.

FlowForge — Multi-Agent Sales Automation | [Github](#)

2026

- LangGraph pipeline of 3 coordinated agents (researcher, analyst, writer) for B2B outreach automation; reduced manual prospecting steps from 6 to 1 by integrating Tavily search, lead scoring, and Gmail dispatch into a single workflow.

MeezanGPT — Corrective RAG Chatbot | [Github](#)

2026

- Islamic banking chatbot using CRAG — a retrieval verification layer that grades context quality before generation, measurably reducing hallucination rate versus standard RAG on domain-specific queries.

ArXiv Lens | [Github](#)

2026

- RAG-based research assistant that ingests academic PDFs, generates structured summaries, and constructs citation-aware knowledge graphs; built with LangChain, ChromaDB, and Groq LLMs.

PerfWatch — VM Attack Detector | [Github](#)

2026

- BiLSTM trained on hardware performance counters to detect VM side-channel attacks in real time; deployed on Hugging Face with a FastAPI backend achieving sub-2s alert latency on rolling log streams.

Full-Stack

- GCRD Student Portal | [URL](#)2025
- Production RBAC student recruitment portal built for UK clients using React, Node.js, MySQL, Docker, and Nginx, with document uploads and multi-stage application tracking, deployed via GitHub Actions CI/CD.
- University Student Hub | [URL](#)2025
- MERN-stack call-centre coordination platform designed for student management workflows, featuring real-time coordination and CI/CD deployment using GitHub Actions.

GPU / Systems & Model Engineering

- Custom CUDA Kernels | [Github](#)2026 – ongoing
- Growing collection of CUDA kernels written from scratch: tiled matrix multiplication with shared memory achieved ~2.8× speedup over naive baseline; fused softmax reduces memory traffic by eliminating intermediate allocations.
- HuggingFace Fine-tuned LLM Suite | [Hugging Face](#)2026
- Fine-tuned and released multiple open-source LLMs (TinyLlama-1.1B, Qwen2.5-7B, OPT-125M, Mistral-7B) using QLoRA with 4-bit quantization; published model cards and reproducible training configs.

ACHIEVEMENTS & CERTIFICATIONS

- [AI Agent Olympics Hackathon](#) — LabLab.ai | Milan AI WeekMay 2026
 - [AI Agents Hack with LabLab & MindsDB](#) — LabLab.ai2026
 - [Meta Hacker Cup](#) — Competitive Programming2024
 - [Introduction to LangGraph](#) — LangChain AcademyMar 2026
 - [LLM Serving & Inference with vLLM and MI300X](#) — AMD AI AcademyApr 2026
 - [Machine Learning Specialization](#) — Stanford / DeepLearning.AI (Andrew Ng)Nov 2025
 - [Data Scientist Associate](#) — DataCampJun 2025