Saad Brohi

Email: Saadbrohi008@gmail.com
Phone: +92-3332197912
LinkedIn: linkedin.com/in/saad-brohi-81521625a
GitHub: github.com/SaadBrohi

Career Objective

Final-year Computer Science student specializing in AI/ML engineering, with hands-on experience in Retrieval-Augmented Generation (RAG), LLM pipelines, computer vision, and production ML deployment. Passionate about building scalable, real-world AI solutions using modern tools like LangChain, FAISS, YOLOv8, Docker, and FastAPI.

Education

National University of Computer and Emerging Sciences (FAST-NUCES) *2022 – Present*
Bachelors in Computer Science

Skills

Programming: Python, C++, C, JavaScript
Backend/APIs: Flask, FastAPI
ML/DL: PyTorch, TensorFlow, Hugging Face Transformers, Diffusers
LLMs/RAG: Phi-3, LLaMA, LangChain, llama-cpp-python, FAISS, SentenceTransformers
Computer Vision: YOLOv8, OpenCV, ByteTrack, Optical Flow
Cloud/DevOps: Git, Docker, Streamlit, AWS, CI/CD (GitHub Actions), SerpAPI
Databases: MySQL, Oracle SQL, MongoDB

Experience

AI Intern – Nexium *June 2025 – Aug 2025*

- Developed end-to-end AI applications using LLMs integrated with LangChain and FAISS.
- Designed RAG-based chat workflows, modeled on QueryVerse, with API support for structured query handling.
- Implemented persona-based memory systems and reasoning using lightweight LLMs.
- Deployed updates via Docker in Agile sprint cycles.

Machine Learning Intern – Elevvo Pathways *Aug 2025 – Sep 2025*

- Built and deployed ML models into **Flask APIs** for real-time inference in production.
- Assisted in developing supervised learning pipelines for predictive analytics using Python and Scikit-learn.
- Conducted preprocessing and feature engineering on structured datasets to improve accuracy.
- Collaborated with senior engineers to test and document ML workflows for deployment.

Projects

QueryVerse – Multi-Document RAG Chatbot *View on GitHub*

- Built a RAG system using **Phi-3-mini**, FAISS, and SentenceTransformers for contextual multi-document querying.
- Designed a Streamlit interface with persona memory, contextual chat, and Markdown rendering.
- Developed and deployed **FastAPI endpoints** to serve RAG-based responses and connect with the UI.

Story2Comic – AI Story-to-Comic Generator *View on GitHub*

- Automated conversion of narrative text to comic panels using LLMs and Stable Diffusion.
- Combined character prompt extraction, scene layout logic, and Hugging Face Diffusers for artwork.
- Delivered full comic output with custom styling, speech bubbles, and sequential storytelling.

Football Match Analysis System *View on GitHub*

- Built a real-time analysis tool to track players and ball using YOLOv8 and ByteTrack.
- Applied optical flow to correct jitter and used KMeans for team identification.

Certifications

- Machine Learning in Production – Coursera (July 2025)
- IBM RAG and Agentic AI – (July 2025)
- Generative AI and LLMs – (July 2025)

Achievements

- Winner – Coders Cup 2025
- Finalist – Coders Cup (2022–2024)
- Final Round - Developers Day (Pseudo Wars) (2023-2024)
- Participant – AI Nexus Competition 2024