SAURABH SINGH

+65-87401149 | saurabh241singh@gmail.com | www.linkedin.com/in/saurabh-singh-0528/ |
www.github.com/saurabh285| www.saurabh285.com

Summary

AI-focused Data Scientist with expertise in LLMs, MLOps, and end-to-end machine learning systems, delivering scalable solutions across NLP, multimodal, and real-time applications.

EDUCATION

National University of Singapore Jan 2022 – Jan 2024
Masters in Computing, Artificial Intelligence specialization (Dissertation Track)

SRM Institute of Science and Technology Aug 2015 – May 2019
Bachelor of Technology in Computer Science and Engineering

Work Experience:

Machine Learning Engineer (Applied Research) May 2024 - Present
Nanyang Technological University Singapore

- Led development of an **LLM-based multimodal system** integrating computer vision and NLP for discourse detection, author classification, and contextual image-text analysis in online group discussions.
- Engineered a sentiment shift forecasting model using user roles, topics, and sentiment trends; achieved **90.91% accuracy**, improving baseline by **15%.**
- Applied end-to-end pipelines (**Airflow, Docker, AWS Sagemaker**), implemented **CI/CD** and **model monitoring** for scalable ML operations.
- Coordinated hybrid **LLM-ML** misinformation pipeline with 6-member team; improved accuracy by **10%** on **1M+ data points** using LangChain and HuggingFace.

AI Assistant – Machine Learning & Computer Vision Jul 2022 – Nov 2023
NUS Human Computer Interaction Lab Singapore

- Developed an advanced eye gaze recognition model utilizing graph neural networks and progressive optimization, achieving **98% precision**.
- Refined preprocessing pipeline and enhanced data processing on eye gaze dataset, achieving **97.67%** accuracy on unseen data, accepted as research thesis at NUS,2024.
- Benchmarked model performance against ANN, CNN, GRU, and LSTM, achieving a **4% improvement** in accuracy.

Software Engineer R&D Jan 2019 - Oct 2021
Hewlett-Packard Enterprise Bengaluru, India

- Spearheaded core module development for HPE's storage platform (SEMC) across **6 major release cycles**, reducing post-deployment bugs by 34%.
- Created **scalable and reusable** UI components using React and Node.js, improving system reliability and UX across enterprise clients.
- Partnered with global product teams in Japan to drive full-stack localization, boosting international user adoption by **25%**.
- **Streamlined CI/CD pipelines** using GitHub Actions and Jenkins; accelerated build-test-deploy cycle times and improved release velocity.
- Collaborated **cross-functionally** with QA, design, and product teams in Agile sprints to align technical execution with business goals.

Projects on GenAI, AI Agents, & ML (www.saurabh285.com):

- **Multi-Domain RAG System:** Built an intelligent multi-domain system by implementing **agentic routing**, **adaptive retrieval**, and **self-reflective reasoning** using DistilBERT, FAISS, LangChain, and Google Gemini
- **Agentic Productivity Assistant** – A personal productivity tasks like email summarization, meeting scheduling, and research retrieval using a multi-agent system with Gemini 1.5 Flash and Google Calendar/Gmail APIs.
- **Latency & Scalability Optimized RAG Architecture** – Engineered a production-ready, low-latency RAG pipeline using FastAPI, Redis, FAISS, and Gemini LLM; achieved **sub-10ms late**ncy via Redis, **~77ms** via FAISS, and **~424ms** with Gemini streaming, optimized with Uvicorn/Gunicorn and Docker.
- **Face Spoofing Detection** – Designed an SVM-based model with feature engineering to detect fraudulent facial attacks on authentication systems; published in IJEAT, 2019.

Technical Skills

- **Languages & Programming:** Python, SQL, JavaScript (ES6+), React, Node.js
- **AI/ML Frameworks & Libraries:** PyTorch, TensorFlow, Keras, Scikit-Learn, Hugging Face Transformers, LangChain, MLflow
- **LLMs & Generative AI:** OpenAI GPT, Google Gemini, RAG Architectures, Prompt Engineering, Transformers, LangChain Agents
- **ML Engineering & MLOps:** Docker, Kubernetes, CI/CD, Jenkins, Airflow, MLflow, Model Deployment & Monitoring
- **Vector Search & Retrieval:** FAISS, Pinecone, ChromaDB
- **Cloud Platforms & Big Data**: AWS (EC2, S3, SageMaker), Azure, Apache Spark, Hadoop, NoSQL (MongoDB, DynamoDB), Grafana
- **Domains & Specialties**: Natural Language Processing (NLP), Misinformation Detection, Sentiment Analysis, Time Series Forecasting, Recommendation Systems, LLM-Based Applications

Other Experiences

Web Development Intern Jun 2017 - Jul 2017

Webnish Software Solutions Bengaluru Area, India

- Delivered a dynamic blogging website using Atom; **Dockerized** the app to achieve faster deployment than comparable projects.
- Deployed and integrated secure user authentication with **Flask-Login**, enhancing personalization and user engagement.
- Executed user input storage using **Azure MySQL**, revamped schema, and queries to reduce response time by 30% and improve data consistency.