Saurabh Chauhan

schau57@uis.edu | +1 (217) 862-4640 |LinkedIn | GitHub | Springfield, IL

PROFESSIONAL SUMMARY

AI Engineer specializing in production generative AI systems and RAG architectures deployed across enterprise environments. Expert in building scalable pipelines using LangChain, foundation models, and vector databases with focus on low-latency inference. Currently pursuing Master of Science in Computer Science with advanced coursework in deep learning and neural networks

TECHNICAL SKILLS

- **Generative AI & Foundation Models:** Microsoft AutoGen, LangChain, RAG, Prompt Engineering, Fine-tuning, OpenAI API, FAISS, Vector Databases, LLMs, NLP, Multi-modal AI
- **ML & Deep Learning:** TensorFlow, Keras, scikit-learn, SpaCy, NLTK, Transformers, BERT, Custom NER
- **MLOps:** Docker, AWS (ECS, EKS, S3, Lambda, SageMaker), GCP, Apache Airflow, PySpark, CI/CD (Jenkins, GitHub Actions)
- **Backend:** Python, Java, SQL, FastAPI, gRPC, PostgreSQL, Redis, Microservices Architecture, Distributed Systems
- **Responsible AI:** Model Evaluation (relevance, bias, hallucination detection), Fairness Assessment

EXPERIENCE

Product Dossier Solutions Pvt Ltd (Kytes) **June 2023 - July 2024**
Software Engineer, AI-ML & Product Development Pune, MH, IN

- Architected and deployed a production **Retrieval-Augmented Generation (RAG)** system leveraging **LangChain**, **Mistral-7B foundation model**, and **FAISS** vector database, implementing **prompt engineering** techniques including few-shot prompting and **context window** optimization to serve **10,000+ users**
- Engineered **MLOps** data processing infrastructure by migrating legacy **Airflow** workflows to a distributed **PySpark architecture** (**Databricks**) with parallel execution patterns, reducing model training data pipeline latency by 65%
- Implemented production safety **guardrails** using **LangChain's** moderation chains and custom validation logic to enforce **responsible AI** controls on **RAG system** outputs

Dasha Krit Technology Pvt Ltd **April 2021 - April 2023**
Software Engineer Pune, MH, IN

- Architected a **multi-tenant** SaaS POSH compliance app using **Django** and **PostgreSQL** by designing custom authentication, 10+ relational models, FSM (transitions), RESTful APIs and deployed it on **GCP** to 10+ clients.
- Streamlined real-time **WebSocket** data pipeline with **SQL Alchemy** processing 1000+ updates/minute, enabling 100+ traders to make data-driven decisions.

University of Illinois Springfield **April 2025 - Present**
Website Intern Springfield, IL

- Developing production-ready web components using **Drupal** CMS, **Twig templates**, **Bootstrap**, and semantic HTML5, ensuring WCAG 2.1-AA accessibility and responsive design across mobile and desktop platforms.

EDUCATION

University of Illinois Springfield **August 2024 - May 2026**
Master of Science, Computer Science Springfield, IL

Pune University **August 2016 - May 2021**
Bachelor of Engineering, Computer Engineering Pune, MH, IN

PROJECTS

- **Multi-Agent Research Assistant System:** Engineered a Microsoft AutoGen multi-agent system (Research, Analysis, Writing) that processes 10+ query types in <5s; integrated Tavily Search with robust error handling to synthesize 8+ sources per request
- **Neurofinity** | Developed AI-powered **mind map generator** using **OpenAI Whisper** for speech-to-text, **Hugging Face Transformers** for key phrase extraction, and **Graphviz** for visualization and reduced note-taking time by 80%.
- **ASL-to-Text: Engineered an ASL-to-text platform** by fine-tuning a **VideoMAE Transformer** on a **239-class dataset**, boosting accuracy from **62% to 82% (+32%)**, implemented **Universal Temporal Sub-sampling** for GPU-constrained training and applied **Responsible AI** principles to reduce recognition variance by **18%** across diverse signing styles
- **Invoice IQ** | Built end-to-end invoice extraction pipeline using custom **SpaCy transformer NER model**, achieving **95% F1-score**, extracting 8 entity types from 200+ invoices and deployed as **Flask API** with GPU-accelerated **Tesseract OCR**

CERTIFICATE

- AWS Machine Learning Essentials (Nov 2025) | AWS Essentials (Jun 2024)