Saurabh Chauhan

schau57@uis.edu · +1 (217) 862-4640 · Chicago, IL
linkedin.com/in/saurabh-chauhan2298 · github.com/saurabh-chauhan22

PROFESSIONAL SUMMARY

AI Engineer with 3+ years of experience specializing in production generative AI systems and RAG architectures deployed across enterprise environments. Expert in building scalable LLM pipelines using LangChain, PyTorch, and vector databases with a focus on low-latency inference, responsible AI guardrails, and multi-agent orchestration. Completing an M.S. in Computer Science at the University of Illinois Springfield (GPA 3.8, expected May 2026) with graduate coursework in deep learning and neural networks.

TECHNICAL SKILLS

Generative AI & Foundation Models: AutoGen, LangChain, LangGraph, RAG, Prompt Engineering, Fine-tuning, OpenAI API, FAISS, Vector Databases, LLMs, NLP, Multi-modal AI
ML & Deep Learning: TensorFlow, PyTorch, Keras, scikit-learn, SpaCy, NLTK, Transformers, BERT, Custom NER
MLOps: MLflow, 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

PROFESSIONAL EXPERIENCE

University of Illinois Springfield · Website Intern · Springfield, IL · Apr 2025 - Present

- Build an OpenAI-powered chatbot with LangChain for contextual FAQ responses, implementing persistent session management and token-limited API calls, handling 200+ queries with 90% accuracy
- 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

Product Dossier Solutions Pvt Ltd (Kytes) · Software Engineer (AI-ML) · Pune, MH, India · Jun 2023 - Jul 2024

- 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 across 6 enterprise clients with 95% query accuracy
- 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%
- Built AI inference APIs using gRPC and integrated with Spring Boot microservices, handling 110,000+ daily requests with intelligent caching and load balancing, achieving 99.5% uptime
- 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 · Software Engineer · Pune, MH, India · Apr 2021 - Apr 2023

- Delivered POSH compliance application to 10+ enterprise clients across HR, legal, and financial sectors, by architecting a multi-tenant SaaS app from scratch using Django, PostgreSQL, FSM-based case management, and RESTful APIs deployed on GCP.
- Enabled 100+ traders to execute data-driven strategies, processing 1,000+ market updates/minute with real-time accuracy, by engineering WebSocket and automated ingestion pipelines in Python and SQLAlchemy to deliver live NIFTY 50 feeds and historical datasets for backtesting and paper trading

PROJECTS

- **Montgomery AI Navigator (World Wide Vibes Hackathon 2026)** - Collaborated with a cross-functional team to build and ship a full-stack civic tech platform in 48 hours using React 18, TypeScript, FastAPI, LangGraph, and Google Gemini, delivering features including services navigation, job matching, community news, and AI chat. Engineered an AI-powered Roadmap Generator using a Retrieve Reason and Validate pipeline with Google Gemini 2.5 Flash, producing personalized step-by-step civic service plans based on citizen income, household size, and location
- **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. Deployed it end-to-end on AWS EKS with React/Vite frontend, FastAPI backend, Docker + Amazon ECR, and an Application Load Balancer across a 2-replica Kubernetes cluster with zero-downtime rolling deployments
- **Neurofinity** - Fine-tuned Flan-T5 (LoRA/PEFT) on meeting transcripts to auto-generate hierarchical mind maps, deploying a FastAPI + Streamlit application with neurodivergent accessibility controls (contrast, spacing, themes) and a custom structural evaluation framework measuring tree edit similarity and hierarchical depth

EDUCATION

Master of Science in Computer Science · University of Illinois Springfield · Springfield, IL, US · May 2026 · GPA: 3.8/4

Bachelor of Engineering in Computer Engineering · Pune University · Pune, MH, India · May 2021

CERTIFICATIONS & LEADERSHIP

AWS Machine Learning Essentials (Nov 2025) · AWS Essentials (Jun 2024) · Secretary, Rotaract Club UIS · IEEE Member