Shubham Jain

+1 (551)-371-2984 | sjain71@stevens.edu | Linkedin | My Website

Work Experience

NxtCRM.AI May 2025 - Aug 2025

AI Engineer Intern (First Intern Hired) *New York, USA*

Tech Stack : Python, Claude AI, OpenAI, GraphRAG, Qdrant, Neo4j, PostgreSQL, Redis, OAuth 2.0, DeBERTa, Vector Embeddings

- Addressed data fragmentation across social platforms by building AI entity extraction with deduplication, processing 10,000+ profiles from LinkedIn/Eventbrite/Facebook/X with high accuracy
- Solved complex query challenges by implementing smart routing with @mention parsing and GraphRAG integration, achieving sub-second contextual responses
- Unified relationship insights by integrating 4 databases with automated timeline extraction, reducing response time60% while enabling relationship strength scoring

Deloitte USI July 2019 – July 2024

Software Engineer *Hyderabad, INDIA*

Tech Stack : Java, Javascript, Python, Selenium, Apache Spark, AWS SageMaker, RESTful APIs, ETL, Computer Vision, ML Models, COFTA, JIRA, Confluence, ALM

- Received Deloitte Spot Award for architecting end-to-end API and UI integration for Disney Seaware Band Integration project, developing RESTful APIs and automated testing framework using Selenium WebDriver and Java, automating 3,000+ test cases with parallel execution, reducing regression testing from 2 weeks to 3 days and saving 2,500 hours annually
- Engineered distributed ETL pipelines using Apache Spark for Walmart's retail data centralization, processing 50+ TB daily from 4,700+ stores with Java-based transformations and real-time data streaming, achieving 70% reduction in processing time and enabling predictive analytics for inventory optimization
- Integrated computer vision capabilities using Amazon SageMaker into Deloitte's proprietary COFTA testing tools, deploying custom ML models for automated UI validation, achieving 85% accuracy and40% reduction in manual QA effort

Research Experience

Accelerating RLHF-based LLM Alignment with Serverless Computing Summer 2025

Summer Research Assistant under Dr. Hao Wang *New Jersey, USA*

- Conducted distributed RLHF experiments on Qwen models using VERL framework, implementing instruction tuning with LoRA adapters across NVIDIA A6000 GPUs and AWS EC2 instances, modifying core framework code to enable serverless computing integration for scalable training workflows
- Making RLHF faster and more cost-efficient while enhancing LLM response quality in reasoning, mathematics, and coding

Developing an autonomous robot with integrated AI decision-making capabilities January 2025-May 2025

Research Assistant under Professor Shucheng Yu *New Jersey, USA*

- Deployed SLAM-based autonomous navigation system achieving 98% path planning success rate on custom robot with mecanum wheels, implementing Python/C++ ROS Noetic architecture that leverages Jetson Orin Nano's GPU (1024 CUDA cores, 67 TOPS) for parallel processing of RPLiDAR point clouds (30,000+ points/sec) with YOLO object detection, dynamic obstacle avoidance, and real-time decision-making at 20+ FPS with 100ms sensor-to-actuator latency

Technical Skills

ML/Data Science: PyTorch, TensorFlow, Keras, OpenCV,Numpy, Pandas, Matplotlib,Jupyter Notebook, Tableau, GitHub, JIRA
Programming Languages: Python, CUDA C, Java, Javascript , SQL, Cypher (Neo4j), Java-Selenium, C, C Sharp, HTML
Cloud & Databases: AWS, GCP, Neo4j, PostgreSQL, Redis, Qdrant, Docker
GenAI Technologies: LangChain, Crew AI, AutoGen 0.4, LangGraph, RAG, KnowledgeGraph,MCP
Certificate: Multi Agent System CrewAI, AWS CCP,Google GenAI,Mastering MCP:Building Advanced Agentic Applications

Projects

Personalized Alumni Engagement Engine Powered by RAG and Multi AI Agents

- Built 5-agent RAG system (Research, Matching, Outreach, Analytics, Coordinator) using AutoGen 0.4 and Chroma vector database, processing 10,000+ alumni profiles with OpenAI embeddings, achieving 92% relevance accuracy
- Deployed on FastAPI/Quart/React stack with WebSocket and LangChain-orchestrated GPT-4/Claude-3, reducing manual effort by 75% and increasing engagement by 3 times

DealFinder AI: Intelligent Multi-Agent Shopping Assistant with NLP

- Developed multi-agent e-commerce system using Gemini LLM for natural language processing and LangChain orchestration, implementing 6 specialized agents (Controller, NLP Parser, Scraper, Aggregator, Comparison, Presentation) that achieved 85% accuracy in finding optimal deals with 3.2-second response time (98% faster than manual search)
- Built BeautifulSoup scraping pipeline for Amazon/Walmart/eBay with async processing, chat memory, and effective price calculations, achieving 98% retailer coverage

Education

Stevens Institute of Technology, Hoboken May 2026

Master of Science in Applied Artificial Intelligence *GPA: 3.945*

Relevant Courses: Machine/Deep Learning, GenAI, GPU and Multicore Programming, Computer Vision, Big Data and Analytics

Jaypee Institute of Information Technology, Noida June 2019

B. Tech in Electronics and communication Engineering

Relevant Courses: Data Structures, Introduction to IOT, Algorithm and Artificial Intelligence and Quantum Mechanics