ABHISHEK SAGAR SANDA

320 Warren St, Roxbury, Boston, MA 02119 | +1 857-395-9451 | sabhisheksagar200@gmail.com | LinkedIn | Github | Portfolio

Summary

Graduate AI engineer specializing in LLM applications and computer vision, with hands-on experience fine-tuning YOLOv8 and GPT-4 for multimodal detection and building RAG pipelines. Proficient in Python, PyTorch, HuggingFace Transformers, LangChain, and prompt engineering. Passionate about developing and deploying scalable applied AI systems to enhance user experiences.

EDUCATION

Northeastern University, Boston, MA **Sep 2023 - Dec 2025**

Master of Science, Information Systems *Boston, MA*

- **GPA:** 3.85
- **Achievements:** Recognized as a top-10 finalist in Murf.AI Coding Challenge, Winner of Northeastern's Roli.AI Hackathon
- **Coursework:** Theory and Practical Applications of Al Generative Modeling, Advanced LLM Techniques, NLP Engineering

PROFESSIONAL EXPERIENCE

Northeastern University **Sep 2025 - Dec 2025**

Teaching Assistant *MA*

- Assisted faculty in facilitating lectures, grading assignments, and guiding students on core machine learning and data science tools and methodologies, leading to improved student comprehension and engagement
- Conducted office hours and provided academic support to over 50 graduate students on advanced generative AI concepts, enhancing their understanding of RNNs, LSTMs, Transformer models, GANs, reinforcement learning, and responsible AI development while demonstrating a passion for learning

Virtual Presenz **Sep 2024 - Dec 2024**

Research Software Engineer Intern *MA*

- Designed retrieval-augmented generation pipelines and fine-tuned YOLOv8 and GPT-4 for multimodal weapon detection and automated report generation, achieving 85% detection accuracy on 70,000+ training images while developing an automated labeling pipeline that processed 10,000+ safety incidents and reduced manual review time by 60%
- Built context-aware chatbot support system for public safety officers using TypeScript, LLM chaining, and memory optimization, collaborating with cross-functional teams to deploy edge-optimized models that reduced response latency by 50% and improved data preparation efficiency by 40%

HCL Technologies **Aug 2022 - Aug 2023**

Graduate Engineer Trainee, Full Stack Development *Chennai, India*

- Delivered secure, scalable .NET and TypeScript enterprise applications using agile methodologies, achieving 90%+ project efficiency while reducing support tickets by 10% and decreasing security vulnerabilities (SAST/DAST issues) by 20% through adaptive back-end optimization

PROJECTS

Al-Powered Interview Coaching IVR System | tel:+18888056555

- Developed a full-stack Al-powered IVR platform using Node.js, Express, PostgreSQL, and Twilio, enabling real-time, voice-based interview simulations and automating feedback for mock interview sessions with secure JWT authentication and RESTful APIs.
- Integrated OpenAI GPT-4 and MurfAl services for intelligent question generation and text-to-speech, optimizing API performance and reducing average response latency by 40%, resulting in a scalable, production-grade system deployed on Railway

RAG-Powered University Chatbot | https://northeastern-university-chatbot.vercel.app/

- Built an end-to-end Al-powered chatbot system for Northeastern University using Python, FastAPI, Scrapy, and ChromaDB, enabling real-time natural language search and Q&A over 80,000 scraped university web pages.
- Engineered a robust data pipeline for large-scale web scraping, semantic document indexing, and retrieval-augmented generation (RAG), with a modern web frontend and automated data management for production readiness.

Al-Powered Richard Wyckoff Trading Assistant

- Built a full-stack Wyckoff Trading Assistant using Flask, PyTorch, and Chart.js, featuring a transformer-based chatbot (8-head, 6-layer model with 1,189 Q&A pairs), real-time market analytics, 6+ technical indicators, and 3 REST APIs with robust error handling and responsive Bootstrap UI.
- Implemented Q-learning reinforcement learning backtesting engine with ε-greedy strategy over 1,000 training episodes, achieving 15% improved ROI prediction accuracy while optimizing performance through lazy loading and caching, reducing API response time by 40% for scalable trading experiments

SKILLS

- **Programming & Frameworks**: Python, JavaScript, Java, C/C++, C#, PyTorch, TensorFlow, OpenCV, ReactJS, NodeJS, Flask, HuggingFace Transformers, LangChain, TypeScript
- **AI/ML Specialization**: LLM fine-tuning, NLP engineering, Computer Vision, OpenAI API, GPT-3.5/4, LLaMA, RAG pipelines, prompt engineering, reinforcement learning
- **Tools & Infrastructure**: CUDA, REST APIs, MongoDB, MySQL, ChromaDB, Git, VSCode, embedded systems (Raspberry Pi, Arduino, NVIDIA Jetson)
- **Core Competencies**: Machine Learning, Passion For Learning