ABHISHEK SAGAR SANDA

Boston, MA | +1 857-395-9451 | sabhisheksagar200@gmail.com | linkedin.com/in/sandaabhisheksagar | abhisheksagarsanda.com

PROFESSIONAL SUMMARY

Senior AI Developer with proven experience leading design, development, and testing of full-stack Python GenAI applications RAG pipelines, document ingestion workflows, vector databases, and agentic AI systems deployed in production. Deep expertise in RAG architecture, document processing pipelines, LangChain/LlamaIndex orchestration, and vector database optimization (ChromaDB, Pinecone, MongoDB Atlas-equivalent). Integrates Agentic AI and MCP technologies into AI solutions; deploys on Azure and GCP (Vertex AI). Builds AI/ML models for data analytics, prediction, and quality forecasting alongside production GenAI systems. Oversees the full software development lifecycle delivering high-quality solutions with minimal supervision communicates effectively with technical and non-technical stakeholders providing clear updates and documentation.

TECHNICAL SKILLS

GenAI & RAG Systems: RAG pipelines · document ingestion pipelines · document processing workflows · vector database optimization · LangChain · LlamaIndex · LangGraph · Agentic AI · MCP (Model Context Protocol) · OpenAI GPT-4 · Anthropic Claude · hybrid retrieval · reranking · confidence scoring · source attribution · grounded generation

Full-Stack Python Development: Python (strong) · FastAPI · Flask (familiar) · REST APIs · Node.js · Express · TypeScript · JavaScript · Next.js · React · async processing · multi-tenant architecture · Celery · Prisma ORM · webhook orchestration

Vector Databases & Data: MongoDB Atlas (familiar) · ChromaDB · Pinecone · PostgreSQL · Redis · semantic search · embeddings · query expansion · structured & unstructured data · document chunking · ETL/ELT pipelines

AI/ML & Analytics: AI/ML model development · data analytics pipelines · prediction models · quality forecasting · confidence scoring · model evaluation frameworks · BLEU/ROUGE (familiar) · LLM evaluation · performance analytics · automated feedback loops

Cloud Platforms: Azure (Azure AI, Azure OpenAI, Azure DevOps) · GCP (Vertex AI, Cloud Run — familiar) · AWS (Lambda, S3, SQS, EventBridge, Bedrock, RDS, IAM) · serverless · cloud-native AI deployment · Docker · Kubernetes (familiar)

DevOps & Engineering: Git · GitHub Actions · CI/CD · unit/integration testing · Agile/Scrum · SDLC · production monitoring · observability instrumentation · GitHub Copilot

PROFESSIONAL EXPERIENCE

Senior AI Developer - G-nee, Boston, MA
Jan 2025 – Mar 2026

- Led design, development, and testing of full-stack Python GenAI application FastAPI backend, React frontend, and AI orchestration layer (LangChain + LangGraph + OpenAI + ChromaDB) overseeing the full software development lifecycle and delivering high-quality production system handling real users 24/7.
- Designed and implemented RAG application in production: document ingestion pipeline (ingestion → chunking → embedding → vector storage → retrieval → context assembly → response validation) with hybrid retrieval combining dense semantic search, keyword matching, and query expansion improving routing accuracy from 78% to 94%.
- Built AI/ML models for analytics and prediction: confidence scoring for output quality forecasting, intent classification models for routing prediction, and QA scoring pipelines for performance analytics providing measurable data-driven insights on AI system behavior.
- Integrated Agentic AI and MCP-compatible tool architecture into AI solution LangGraph multi-agent orchestration with deterministic routing for triage decisions and model-driven generation for response synthesis.
- Communicated effectively with technical and non-technical stakeholders translated agentic architecture and AI performance metrics into clear business impact documentation and progress updates without jargon.

Research Software Engineer Intern - Virtual Presenz Inc., Shrewsbury, MA
Sep 2024 – Dec 2024

- Built and deployed full-stack Python AI application (YOLOv8 + GPT-4) for public-safety analytics designed document processing workflows and REST APIs with observability instrumentation for downstream system integration.
- Co-authored research on AI/ML model optimization (TensorRT quantization, FP16/INT8 benchmarking, performance prediction); presented at MIT IMPACT Symposium 2025.

.NET Full Stack Developer - HCL Technologies, Chennai, India
Aug 2022 – Aug 2023

- Delivered full-stack enterprise applications (C#, .NET, SQL Server) in Agile environments with CI/CD via Azure DevOps strong SDLC foundations for full-stack Python development and cloud-native AI deployment.

AI PROJECTS

Northeastern University AI Chatbot - Production RAG Platform
2025 – Present

- Designed and implemented production RAG application full document ingestion pipeline and processing workflow (80,000+ documents), vector database optimization (ChromaDB, equivalent to MongoDB Atlas vector search patterns), hybrid retrieval, and grounded response generation. 99.7% uptime, sub-5s latency, 50+ concurrent queries.

- Built AI/ML quality forecasting model: confidence scoring predicts response quality before delivery, groundedness validation detects hallucination risk, and automated regression monitoring provides performance analytics measurable data-driven quality control.
- Led full software development lifecycle independently: requirements gathering, architecture design, development, testing (automated CI/CD), deployment (AWS Lambda, S3), and ongoing monitoring high-quality delivery with minimal supervision.

SupportCopilot - Multi-Tenant GenAI Platform *2025 – Present*

- Led design and development of full-stack Python GenAI application (FastAPI + PostgreSQL + Redis + ChromaDB) with document ingestion pipelines, vector retrieval orchestration, Agentic AI workflows, and MCP-compatible tool architecture across isolated multi-tenant enterprise environments.
- Deployed on AWS (S3, SQS, EventBridge, Lambda) with CI/CD automation (GitHub Actions, 92% test coverage) managed full SDLC from architecture through production deployment and monitoring with clear stakeholder communication.
- Integrated Azure OpenAI-equivalent architecture alongside OpenAI GPT-4 and Anthropic Claude evaluated cloud AI platform performance for production GenAI workloads.

BitVoice - Voice AI + Lightning Network *MIT Bitcoin Hackathon 2026 - 2nd Place | Live: +1 (888) 805-6555*

- Designed and shipped full-stack Python application in 36 hours FastAPI backend, PostgreSQL data layer, OpenAI LLM pipeline, real-time analytics dashboard (SSE + Chart.js), and Railway cloud deployment. Live at web-production-51d43.up.railway.app/dashboard.

AI Interview Coaching IVR *2025*

- Built Node.js + OpenAI voice AI application with document processing (Q&A dataset ingestion), prediction models for question routing, and sub-3s response times under concurrent load live in production.

TEACHING & TECHNICAL LEADERSHIP

Teaching Assistant - CSYE 7380: Theory and Practical Applications of AI Generative Modeling, Northeastern University *Sep 2025 – Dec 2025*

- Mentored 50+ graduate students on RAG pipelines, agentic AI, document processing workflows, and production GenAI deployment developed strong ability to communicate complex AI concepts clearly to diverse technical and non-technical audiences.

AWARDS

- 2nd Place - MIT Bitcoin Hackathon 2026 (BitVoice: full-stack voice AI shipped in 36 hours)
- Winner - Northeastern Roli.AI Hackathon (adaptive conversational AI with LangGraph)
- Outstanding Master's Student Award in Community Impact - Northeastern University (2026)

PUBLICATIONS & PRESENTATIONS

- Co-Author, Real-Time Performance Analysis of TensorRT Optimization for High-Speed Object Detection (2025)
- Poster Presenter, MIT IMPACT Symposium - Virtual Presenz Public-Safety Analytics Work (2025)

EDUCATION

Northeastern University - M.S. Information Systems *Sep 2023 – Dec 2025*

Boston, MA | GPA: 3.82/4.0 | Outstanding Master's Student Award in Community Impact (2026)

Relevant Coursework: AI Generative Modeling, Advanced Techniques with LLMs, NLP Technologies, Cloud Computing

Mahatma Gandhi Institute of Technology - B.Tech, ECE *Aug 2018 – Apr 2022*

Hyderabad, India