NAVEENA POKALA

984-326-1234 • n_pokala@uncg.edu • linkedin.com/in/naveenapokala • github.com/NAVEENAPOKALA

SUMMARY

Innovative AI/ML Engineer with around 6 years of experience researching, building, and deploying production-scale AI systems across cloud and on-premises environments. Deep expertise in Large Language Models, Generative AI frameworks (OpenAI, LangChain, RAG), deep learning (PyTorch), and natural language processing. Proven track record designing high-impact ML solutions with comprehensive MLOps pipelines, model monitoring, and continuous improvement workflows. Passionate about leveraging cutting-edge AI to solve challenging computational problems and drive meaningful business impact.

EDUCATION

University of North Carolina at Greensboro August 2024 – May 2026
Master of Science in Computer Science (GPA: 3.96)

PUBLICATIONS

ACDSA 2026 – Stock Prediction Based on Annual Reports Using Large Language Models

TECHNICAL SKILLS

Languages: Python (3.8–3.12), Java, JavaScript, TypeScript, SQL, Bash, Node.js
Python Libraries: Pandas, PyTorch, TensorFlow, scikit-learn, Transformers, LangChain
ML & AI: Natural Language Processing (NLP), Large Language Models, LangChain, Transformers, Prompt Engineering, RAG, Agentic AI, Text Classification, Sentiment Analysis
Cloud & DevOps: AWS (EC2, Lambda, S3, IAM, CloudWatch), Terraform, RBAC, Docker, Kubernetes, Jenkins, CI/CD, Git, SDLC
Frameworks & APIs: FastAPI, REST APIs, Spring Boot
Databases: MySQL, DB2, Sybase, FAISS Vector DB
Testing: TDD, BDD, Unit Testing (JUnit, Mockito), Integration Testing
Messaging Queues: Kafka, RabbitMQ, ElasticSearch

EXPERIENCE

University of North Carolina at Greensboro, NC — AI/ML Research Assistant February 2025 – Present

• Built end-to-end training and evaluation pipelines for LLM-based models, integrating LLaMA-2 embeddings with CNN architectures to process 10,000+ natural language documents.
• Engineered time-series ML modeling framework processing multi-modal temporal data, integrating textual information with numerical market signals for predictive analytics.
• Developed scalable Python ML pipelines for data ingestion, preprocessing, embedding generation, model training, and evaluation across multiple architectures.
• Engineered retrieval systems using vector embeddings and contextual chunking to align long-form NLP text with temporal events.
• Built distributed training pipelines using PyTorch for training models on large-scale datasets with optimized GPU utilization and batch processing.
• Improved model performance by 32–45% (lower MAPE) versus baseline models by combining LLM representations with time-series modeling.
• Designed and validated novel financial features by discovering statistical patterns in complex, noisy datasets using mathematical modeling and hypothesis testing.

AI-Powered Academic Assistant — GenAI + Agentic System February 2025

• Developed a multi-agent LLM workflow using LangChain, enabling agents to retrieve, summarize, and validate complex academic content.
• Implemented a RAG pipeline using FAISS, HuggingFace embeddings, and Zephyr-7B for context-aware Q&A over thousands of documents.
• Built and deployed a FastAPI backend supporting parallel LLM inference, caching, and latency-optimized retrieval.

• Applied advanced prompt engineering and optimization techniques to improve model prediction accuracy by 35%.
• Implemented production-grade system using PyTorch and HuggingFace with optimized inference latency (less than 200ms) through performance profiling and optimization.

Ernst & Young (Client: Morgan Stanley) —Software Engineer June 2022 – April 2024

• Architected and developed high availability microservices using Java/Spring Boot and Node.js, serving 10,000+ daily transactions with 99.9% uptime across cloud infrastructure.
• Built responsive frontend applications using Angular and TypeScript with modern design patterns, improving user experience and development efficiency by 40%.
• Established comprehensive CI/CD pipelines using Jenkins and Git with automated builds, testing, and multi-environment deployments across AWS cloud platform.
• Implemented event-driven architecture using Kafka and RabbitMQ for asynchronous message processing, improving system decoupling and scalability across microservices.
• Containerized applications using Docker and orchestrated deployments on AWS ECS/Fargate (PaaS), demonstrating hands-on experience with cloud-native architecture.
• Migrated 100+ batch jobs to AWS serverless architecture using Lambda and CloudWatch, reducing operational costs by 25% and improving system reliability.
• Provided L1/L2 production support, proactively diagnosing and resolving live incidents while maintaining SLA compliance and 24/7 system performance.

Tata Consultancy Services (Client: TCS BaNCS) — System Engineer / Team Lead June 2019 – June 2022

• Delivered 50+ full-stack web applications using Java, Spring Boot, Python, and Angular through all phases of project lifecycle including requirements, design, development, and testing.
• Built cloud-native applications with microservices architecture, implementing RESTful APIs and integrating with various backend services and databases.
• Developed comprehensive testing strategies using JUnit, Mockito, and integration testing frameworks following TDD/BDD methodologies, achieving 90%+ code coverage.
• Led Agile development team of 5 engineers, facilitating sprint ceremonies and delivering iterative solutions based on proven Agile methodology.
• Optimized database performance through SQL query tuning and NoSQL data modeling, reducing execution time by 30% and improving application responsiveness.
• Implemented modern collaborative development workflows using Git, code reviews, and continuous integration tools, ensuring code quality and team productivity.

AWARDS & CERTIFICATIONS

• Star of the Month Award – Tata Consultancy Services, Sep 2021
• AWS Certified Cloud Practitioner (CLF-C01) – Valid through Apr 2026