Neha Thokala
thokalaneha3@gmail.com | 817-695-5714 | Houston, TX

Professional Summary

Innovative Generative AI Engineer with 5+ years of experience in building and deploying enterprise-grade AI/ML and LLM-powered applications that drive measurable business outcomes. Specialized in Retrieval-Augmented Generation (RAG), agentic AI workflows, and full-stack Python development on AWS and Azure. Proven ability to lead end-to-end solution delivery —from client requirements to production deployment— blending data engineering, AI architecture, and forward-deployed implementation skills.

SKILLS

- **Programming:** HTML, CSS, JavaScript, jQuery, ASP.NET, Python, C#, Citrix & Virtual Environment Automation, AI Model Development, Machine Learning, Microsoft Power Automate, Power Apps, Power BI, Forms, SharePoint.
- **Generative AI & LLMs:** Prompt Engineering | RAG Pipelines | Context Engineering | Fine-Tuning | RLHF
- **AI Agents & Frameworks:** LangChain | Semantic Kernel | LlamaIndex | Agent Tool Orchestration
- **Vector Databases:** Pinecone | FAISS | Milvus | ChromaDB
- **Cloud & DevOps:** AWS (S3, EC2, Lambda, SageMaker) | Azure | Docker | Kubernetes | CI/CD | Terraform
- **Machine Learning:** TensorFlow | PyTorch | scikit-learn | Model Training | Optimization | Quantization
- **Data Engineering:** ETL (Pandas, Spark, Glue) | SQL | Data Pipelines | API Integration | Big Data Processing
- **Visualization & Analytics:** Looker | Power BI | Matplotlib | Plotly
- **Programming Languages:** Python | JavaScript | SQL | Shell

EXPERIENCE

AI Automation Developer| West Virginia University Medicine **Feb 2025 - Current**

- Architected and deployed GenAI solutions leveraging LangChain, LangGraph, and LlamaIndex, integrating LLM agents for document summarization, context retrieval, and automated decision workflows.
- Designed Retrieval-Augmented Generation (RAG) pipelines with vector databases (ChromaDB, Pinecone) using embedding models to improve context accuracy by 35%.
- Developed Python-based APIs for model serving, retrieval logic, and metadata search within microservice architecture hosted on Azure Kubernetes Service (AKS).
- Fine-tuned OpenAI GPT-4 and Claude 3 models for custom domain knowledge tasks; implemented prompt optimization using LangChain memory and token cost reduction strategies.
- Created cloud-native observability dashboards for model performance tracking and latency monitoring via GCP Cloud Monitoring and Prometheus. Automated AI workflow deployment using Azure DevOps CI/CD, Terraform, and containerization with Docker.
- Designed and implemented Retrieval-Augmented Generation (RAG) pipelines connecting structured and unstructured enterprise data to LLMs (OpenAI GPT-4, Claude 3, Gemini).
- Built and deployed AI agents using LangChain and LlamaIndex for automated knowledge retrieval, report generation, and decision support. Conducted model fine-tuning and RLHF on domain-specific datasets to enhance response fidelity and compliance.
- Developed end-to-end AI-enabled applications with FastAPI backend and react front-end, integrating LLM APIs through secure token management. Implemented vector databases (Pinecone, ChromaDB) to improve semantic search accuracy by 40%.
- Deployed containerized AI services on AWS ECS and Azure Kubernetes Service (AKS) with automated CI/CD pipelines.
- Created LLM observability dashboards tracking latency, accuracy, and token usage via Prometheus and CloudWatch.
- Led client demos, technical workshops, and enablement sessions to drive adoption of AI features and trust in production deployments.
- Spearheaded LiDAR + AI project with Purdue University to integrate iForester tree measurement system into ArborTrue mobile app.

AI Automation Specialist | Capital One **Dec 2022 – Jan 2025**

- Containerized applications using Docker and deployed to GCP Cloud Run and Kubernetes clusters.
- Collaborated with data science teams to integrate NLP models into production using GCP Vertex AI.
- Maintained code versioning through Git/Azure DevOps, adhering to CI/CD best practices and Agile methodologies.
- Designed SQL queries and stored procedures to extract, transform, and load (ETL) data for process automation.
- Deployed bots for financial reporting, invoice processing, and client onboarding, achieving 95% process accuracy.
- Collaborated with cross-functional teams in Agile sprints to deliver automation features on time. Developed SOAP and REST APIs for cross-platform integrations with ERP and supply chain systems, enabling seamless data sharing.
- Automated system tasks and workflows (data imports, validation, and reconciliation) through Python scripts and SQL stored procedures.
- Integrated OpenAI solutions to enhance automation capabilities, enabling intelligent decision-making and task handling.
- Identified automation opportunities across deployment, data management, and service management processes, improving operational efficiency. Wrote and maintained Python and JavaScript scripts for automation and testing.
- Collaborated with enterprise clients to define AI/ML integration requirements and translate business goals into technical solutions.
- Built ETL pipelines aggregating multi-source datasets into AWS S3 data lakes using Spark and Python.
- Deployed machine-learning models on SageMaker and integrated them via REST APIs for real-time inference.
- Supported big-data processing with Hadoop and Spark for analytics workflows serving millions of records.
- Implemented microservices and serverless components using AWS Lambda and API Gateway for event-driven architecture.
- Ensured data integrity through SQL schema design, indexing, and transactional validation for ML pipelines.
- Conducted rigorous testing of automated workflows to ensure reliability, achieving 99% uptime.

- Successfully automated document-heavy processes using OCR and UiPath Document Understanding Module, reducing manual effort by 60% and increasing processing speed. Built automation scripts in Python & VB Script to extend Power Automate bot capabilities.
- Mentored junior developers on UiPath Studio workflows, reusable components, and automation best practices.
- Developed robots using Power Automate in identifying and debugging the errors using error handlers.

Software Developer |DXC Technology **Jun 2021 - Jan 2022**

- Designed and developed microservices using Java 8, Spring Boot and Hibernate ORM.
- Supported development of AI prototypes for business automation.
- Built LLM-driven automation systems for content generation, report summarization, and task orchestration using LangChain and OpenAI APIs. Created domain-adapted LLMs via supervised fine-tuning and quantization for resource-constrained environments.
- Developed AI workflow agents combining planning and execution roles for multi-step reasoning tasks.
- Implemented observability frameworks capturing bias, cost, and accuracy metrics across inference runs.
- Integrated context engineering layers linking knowledge bases, memory stores, and tool functions for adaptive prompt composition.
- Designed serverless pipelines integrating Azure Functions, Google Cloud Storage, and Power Automate to reduce manual intervention.
- Experimented with LLM fine-tuning and prompt-chaining workflows for structured reasoning across multiple agents.
- Developed and managed REST APIs, improving system interoperability and reducing latency by 25%.
- Enhanced database performance through MYSQL query optimization, reducing average query time by 40%.
- Enhanced automation accuracy by integrating ML models for anomaly detection in process workflows.
- Participated in Agile ceremonies and provided technical presentations to non-technical stakeholders.
- Established CI/CD pipelines in Jenkins to automate build, test, and deploy processes, reducing deployment times by 50%.
- Followed the full software development life cycle (SDLC) and agile processes, contributing to sprint planning and daily stand-ups.
- Applied Core Java multithreading to handle concurrent user requests and reduce processing time, achieving real-time notifications.
- Create responsive and visually appealing UI components using HTML, CSS, and JavaScript frameworks like React.

IT Automation Developer |Capgemini **Jun 2019 - May 2021**

- Utilized Python scripts for data manipulation and automation tasks, improving process efficiencies and reducing manual intervention.
- Extensive knowledge of RPA principles and best practices, including exception handling, logging, and transaction processing to optimize reliability and maintainability.
- Developed automation scripts in UiPath and Python to integrate with legacy systems, relational databases, and third-party APIs.
- Experienced in C#, VB.NET, and Python, enabling custom code integration within UiPath to address complex automation requirements.
- Familiar with Process Mining techniques to identify automation potential, streamline processes, and enhance decision-making.
- Computed on Web Applications, Desktop Applications and Windows Applications by using basic, desktop, web recorders and screen scraping & data scraping. Developed automation scripts for quality assurance testing, reducing manual testing efforts by 50%.
- Optimized IT operations by identifying and automating repetitive tasks, saving 300+ hours annually.
- Collaborated with cross-functional teams to define and implement tailored automation solutions.
- Performed the practical usage of various UiPath Orchestrator functionalities - Bots, Processes, Assets, Jobs, Schedulers, Logging, Recovery Methods and Application Credentials.
- Maintained both Attended and Unattended Robot resources, and provided centralized Robot logs, remote execution, monitoring, scheduling, and work queues using UiPath Orchestrator.

PROJECTS

Customer Feedback Analysis and Sentiment Classification

- Use UiPath and ML to extract and analyze customer feedback data from various sources (emails, chat logs). Apply sentiment analysis to classify feedback as positive, neutral, or negative.
- Automate the collection and classification of customer feedback to identify service improvement areas, and generate reports based on sentiment trends. Provide Action Center integration for employees to review and adjust classifications, helping to refine model accuracy over time.

Executive AI Assistant: Built an LLM-based assistant that automated scheduling, meeting notes, and email drafting for startup executives.

Tree Measurement AI: Developed AI + LiDAR pipeline for automated forestry measurements, integrating real-time image, depth, and camera parameter analysis with Purdue's iForester system.

RAG-Powered Knowledge Assistant: Developed an internal assistant using LangGraph + ChromaDB to query large document sets via embeddings and LLM reasoning. Integrated Azure OpenAI API for contextual Q&A and FastAPI backend with React front-end dashboard.

AI Workflow Orchestrator (LangChain + Azure)

Designed a modular AI orchestration framework connecting multiple agents for text classification, data summarization, and report generation. Deployed on Azure Kubernetes Service with Dockerized endpoints and Terraform provisioning.

EDUCATION

Master of Science, *Computer Science*| Arlington, USA| University of Texas at Arlington

Bachelor of Technology, *Computer Science* | Warangal, India| SR University

CERTIFICATIONS

- Microsoft Certified: Azure AI Engineer Associate
- Google Cloud Professional Data Engineer (in progress)