NIVITH AVULA
Generative AI Engineer
USA | +1 (469) 605-3132 | Nivith1029@gmail.com | LinkedIn | Portfolio

PROFESSIONAL SUMMARY

Generative AI Engineer with expertise in architecting and deploying enterprise scale LLM, RAG, and agent based systems across Azure and AWS environments. Proven track record of building production grade AI platforms serving millions of users in supply chain, banking, healthcare, and energy domains. Specialized in fine-tuning LLMs, optimizing retrieval pipelines, and delivering secure, compliant, cloud native AI microservices. Passionate about transforming complex business challenges into scalable, intelligent systems that drive measurable impact and operational excellence.

TECHNICAL SKILLS

Programming Languages: Python, Go, SQL, Bash
Generative AI & LLMs: Azure OpenAI, Amazon Bedrock, OpenAI API, Hugging Face Transformers, LangChain, LangGraph, CrewAI, Ollama, Prompt Engineering, Few-shot Learning, Multi-Agent Systems, Agent to Agent (A2A) Orchestration, RAG (Retrieval-Augmented Generation), LoRA, QLoRA, Fine-Tuning, Token Optimization, LLM Evaluation, Hallucination Mitigation
Machine Learning & Deep Learning: TensorFlow, PyTorch, XGBoost, Scikit-learn, CNN, RNN, LSTM, GANs, Transformers, Time-Series Forecasting (ARIMA, Prophet), Feature Engineering, Model Benchmarking, Bias Evaluation
Vector Databases & Search: Azure Cognitive Search (Vector Search), Amazon OpenSearch, FAISS, Pinecone, Weaviate, Semantic Search, Embeddings
Cloud Platforms: Microsoft Azure (Azure OpenAI, AKS, ACR, Azure Functions, Data Factory, App Service, Blob Storage, Entra ID, Key Vault, API Management), Amazon Web Services (SageMaker, Bedrock, EC2, ECS, EKS, ECR, Lambda, S3, Glue, Kinesis, Textract, CloudWatch, CloudTrail), Google Cloud Platform (Vertex AI, GKE, BigQuery, Dataflow, Cloud Storage)
Data Engineering & Big Data: Databricks, Apache Spark (EMR), Snowflake, Kafka, Azure Event Hubs, Service Bus, Apache Airflow, ETL/ELT Pipelines, Data Transformation, Data Validation
Backend & API Development: FastAPI, Flask, Django, RESTful APIs, Microservices Architecture
Frontend Technologies: React, Next.js, JavaScript, TypeScript, HTML5, CSS
MLOps & DevOps: Docker, Kubernetes, Helm, Terraform, GitHub Actions, Jenkins, Azure DevOps, CI/CD Pipelines, MLflow, SageMaker Pipelines, Kubeflow
Databases: PostgreSQL, MySQL, Cassandra, Neo4j, BigQuery
Monitoring & Governance: Prometheus, Grafana, OpenSearch Dashboards, CloudWatch, Audit Logging, Model Risk Governance, Human-in-the-Loop Validation, UAT
Security & Identity: OAuth2, RBAC, Azure Entra ID, Managed Identities, Private Endpoints, Secure API Design

PROFESSIONAL EXPERIENCE

Blue Yonder | Generative AI Engineer | Coppell, Texas, USA | July 2025 – Present

- Lead the architecture and enterprise deployment of Generative AI solutions for supply chain planning and forecasting using Azure OpenAI, LangChain, LangGraph, and Python, enabling AI driven decision intelligence across mission critical operations.
- Architect multiagent and Agent to Agent (A2A) orchestration frameworks using LangGraph to coordinate retrieval, reasoning, and validation agents for complex, multi step supply chain decision workflows.
- Design and implement scalable Retrieval Augmented Generation (RAG) platforms leveraging Azure Cognitive Search (vector indexing), Azure OpenAI embeddings, Azure Blob Storage, and semantic chunking to power contextual enterprise knowledge retrieval.
- Apply parameter efficient fine tuning techniques (LoRA, QLoRA) using Hugging Face and PyTorch to adapt pretrained LLMs to domain specific supply chain terminology while optimizing compute utilization.
- Develop and govern production grade AI microservices using Python, Go, FastAPI, Azure Functions, and Azure App Service, exposing secure REST APIs consumed by enterprise planning and analytics systems.
- Evaluate and implement vector database strategies (Pinecone, Weaviate) to optimize semantic search performance, relevance scoring, and latency at enterprise scale.
- Establish enterprise data ingestion and transformation pipelines using Azure Data Factory, Databricks (PySpark), Pandas, and Blob Storage to convert ERP and unstructured documentation into vectorized knowledge assets.
- Enhance model reliability and response accuracy by reducing hallucinations by 38% through structured prompt engineering, few-shot design, token optimization, and validation pipelines.
- Enforce enterprise grade security through OAuth2, Azure Entra ID, Managed Identities, Azure API Management, Key Vault, private endpoints, and RBAC controls.
- Operationalize GenAI workloads using Docker, AKS, ACR, Helm, GitHub Actions, and CI/CD pipelines aligned with governance standards, with monitoring via Prometheus and Grafana.

UBS | AI Engineer | New York, USA | July 2024 – June 2025

- Directed the architecture and deployment of enterprise AI and Generative AI platforms for banking operations using Amazon SageMaker, Bedrock, LangChain, LangGraph, and Hugging Face within regulated financial environments.
- Designed multi-agent orchestration frameworks to coordinate policy retrieval, compliance reasoning, and response validation for complex, audit-sensitive workflows.
- Implemented parameter efficient fine tuning (LoRA, QLoRA) to specialize LLMs for financial and regulatory language while maintaining cost and governance controls.
- Engineered secure Retrieval Augmented Generation architectures using Amazon OpenSearch (vector search), FAISS, S3, and embedding models to enable contextual search across enterprise policy repositories.
- Developed high-throughput AI microservices using Python, FastAPI, Flask, Docker, and AWS Lambda, supporting millions of monthly inference requests with sub-second latency.
- Integrated Neo4j knowledge graphs to enhance entity-aware retrieval and context reasoning within compliance intelligence systems.
- Built document intelligence pipelines using Amazon Textract and NLP preprocessing workflows, reducing document processing time by 48%.
- Governed AI workloads using EC2, ECS, EKS, ECR, Terraform, and CI/CD pipelines aligned with regulatory standards and change management frameworks.
- Implemented comprehensive monitoring, audit logging, and compliance observability using CloudWatch, CloudTrail, and OpenSearch Dashboards.
- Ensured model risk governance through benchmarking, bias evaluation, human in the loop validation, and UAT cycles in partnership with risk and compliance stakeholders.

Landis Gyr | Full Stack AI Engineer | Atlanta, GA | Jan 2024 – May 2024

- Delivered end to end AI and Generative AI solutions for smart energy analytics platforms, integrating LLM-driven insights, predictive modeling, and real-time IoT processing to support grid operations.

- Designed autonomous agent based workflows using LangChain and CrewAI to orchestrate anomaly detection, contextual data retrieval, and operational insight generation.
- Developed LLM enabled applications using Amazon Bedrock and Hugging Face within RAG architectures to enable natural-language querying of smart-meter, outage, and grid datasets.
- Built forecasting and anomaly detection models using SageMaker, XGBoost, TensorFlow, and time-series feature engineering, improving forecast accuracy by 14% over baseline models.
- Engineered scalable ingestion and processing pipelines using AWS IoT Core, Kinesis, Glue, S3, Spark (EMR), and OpenSearch to support both batch and real-time ML workloads.
- Operationalized AI models using MLOps best practices including SageMaker Pipelines, experiment tracking, automated retraining, and environment promotion.
- Deployed containerized AI services across Kubernetes environments (EKS/ECS/EC2) with CI/CD automation and production-grade monitoring frameworks.

UHG | Python Full Stack Engineer | Dallas, Texas | Sept 2023 – Dec 2023
- Developed scalable healthcare web applications using Python (Flask, Django, FastAPI) and built responsive user interfaces with React, Next.js, JavaScript, TypeScript, HTML, and CSS, ensuring seamless frontend-backend integration.
- Designed and implemented RESTful APIs and integrated them with React-based dashboards to enable real-time clinical data visualization, patient monitoring, and analytics reporting.
- Engineered machine learning and deep learning models including CNNs, RNN/LSTMs, GANs, and Transformer-based NLP architectures to generate actionable healthcare insights from structured and unstructured datasets.
- Built and deployed end-to-end ML pipelines using Kubeflow on Google Kubernetes Engine (GKE) and managed scalable model serving through Google Vertex AI.
- Developed computer vision and NLP solutions using OpenCV, YOLO, and Hugging Face (BERT, GPT), integrating model outputs into user-facing applications to enhance diagnostic and patient communication workflows.
- Leveraged Google BigQuery, Cloud Storage, and Dataflow for large-scale data processing, and implemented MLflow for model lifecycle management, experiment tracking, and reproducibility.

Celanese | Python Developer | Irving, Texas | May 2023 – Aug 2023
- Designed and developed enterprise microservices using Python (Flask, Django, FastAPI) to support scalable, distributed backend systems.
- Deployed high availability services using Azure Functions and AKS, implementing containerized and serverless architectures.
- Architected event driven systems using Kafka, Azure Event Hubs, and Service Bus to enable real time, asynchronous data processing.
- Built and automated CI/CD pipelines using Azure DevOps and GitHub Actions, implementing infrastructure provisioning and application deployments in multi-cloud environments using Terraform and Ansible.
- Optimized database performance across PostgreSQL, MySQL, Cassandra, and Azure Cognitive Search through indexing and schema improvements.
- Integrated AI services and analytics components into Azure-hosted enterprise applications.

Altimetrik | Python Developer | Pune, India | August 2020 – July 2022
- Designed and developed scalable ETL pipelines using Python and PySpark to process large-scale enterprise datasets, improving reporting efficiency by 50%.
- Built and optimized data workflows using Apache Airflow with robust scheduling, dependency management, and failure handling to ensure reliable pipeline execution.
- Migrated legacy on premise data systems to Azure based Snowflake data warehouse architecture, improving scalability and reducing report generation time.
- Performed complex data transformations, cleansing, validation, and deduplication to enhance data quality and ensure accurate business intelligence reporting.
- Developed RESTful APIs using FastAPI and Flask to expose processed data to downstream applications and analytics platforms.
- Refactored monolithic components into modular, scalable Python based microservices, improving system maintainability and deployment efficiency.
- Containerized applications using Docker and deployed them on Azure Kubernetes Service (AKS) to support scalable, cloud-native architecture.
- Optimized SQL queries and implemented indexing strategies to improve database performance and reduce API response times by up to 30%.
- Integrated CI/CD pipelines using Git and Jenkins to automate build, testing, and deployment workflows.
- Implemented structured logging, monitoring, and exception handling mechanisms to enhance production stability and reduce downtime.

EDUCATION

University of North Texas, Denton, TX
Master of Science in Information Systems and Technology | August 2022 – May 2024

PROJECTS

I**MDB Conversational Voice Agent |** LangGraph, Whisper, OpenAI TTS
Built a production-grade conversational voice agent enabling natural language exploration of the IMDB Top 1000 dataset.
- Designed a LangGraph-based agent workflow to classify user queries into structured SQL queries, semantic vector search, and hybrid retrieval strategies.
- Implemented voice input using Whisper STT and voice output using OpenAI TTS to enable seamless voice-based interaction.
- Integrated semantic search with vector embeddings to enhance contextual movie discovery and recommendations.
- Implemented clarifying follow-up prompts and intelligent recommendations to improve conversational accuracy and user engagement.

Enterprise RAG PDF Assistant | Ollama, FastAPI, FAISS
- Built a local Retrieval-Augmented Generation (RAG) system using Ollama-hosted LLMs to enable secure, cost-efficient document intelligence without reliance on external APIs.
- Designed embedding pipelines and FAISS-based vector indexing to support semantic search and citation-based responses over uploaded PDFs.
- Developed a FastAPI backend to manage document ingestion, chunking, embedding generation, and query handling.
- Implemented citation-aware answer generation to improve trust and explainability in responses. Optimized local inference performance and memory usage for efficient on-prem deployment.

CERTIFICATIONS

- AWS Certified Solutions Architect
- HashiCorp Terraform Associate
- Microsoft Azure Fundamentals (AZ-900)