NIVITH AVULA
Generative AI Engineer
+1(469)-605-3132 | Nivith1029@gmail.com |www.linkedin.com/in/nivith-avula-0b44a0140 | https://nivith-portfolio.vercel.app/

Professional Summary

Generative AI Engineer with **5+ years of experience** building and deploying production-grade AI systems across **SaaS, banking, healthcare, energy, and manufacturing** domains. Strong expertise in **Python and Go backend engineering**, cloud-native architectures, and **LLM-powered applications**, including **RAG pipelines, agentic workflows, and vector search systems**. Proven track record of delivering **secure, scalable GenAI platforms** on **Azure, AWS, and GCP**, integrating pretrained LLMs, MLOps pipelines, and enterprise APIs. Experienced in collaborating with product, UX, and data teams in Agile environments to translate complex business problems into reliable, production ready AI solutions.

Skills

- **Programming & Backend:** Python, Go , Java, Scala, FastAPI, Flask, Django, Spring Boot, REST APIs, gRPC, GraphQL, WebSockets, SQL, NoSQL, SQLAlchemy
- **Frontend & UI:** HTML, CSS, JavaScript, Typescript, jQuery, AJAX, React.js, Node.js, Tailwind, Streamlit, Dialogflow, Haystack, Human-loop
- **Generative AI & LLMs:** Azure OpenAI,Amazon Bedrock, OpenAI APIs, Hugging Face Transformers, PyTorch, LangChain, LangGraph,LlamaIndex, CrewAI Agentic AI, A2A Workflows, LoRA, QLoRA, Prompt Engineering
- **RAG & Vector Databases:** Azure Cognitive Search, Amazon OpenSearch, FAISS, Pinecone, Weaviate, Neo4j
- **Machine Learning & Deep Learning:** TensorFlow, PyTorch, Scikit-learn, XGBoost, CNNs, RNNs, LSTMs, GANs, Transformers, ARIMA, Prophet
- **NLP & Computer Vision:** spaCy, NLTK, BERT, GPT, OpenCV, YOLO, ImageNet
- **MLOps & Model Lifecycle:** SageMaker Pipelines, Kubeflow, MLflow, LakeFS, GitLFS, pachyderm, NeptuneAI, ClearML, Comet ML
- **Cloud Platforms:** Azure (Azure OpenAI, Azure Functions, AKS, ACR, Azure Data Factory, Azure Databricks, Azure Blob Storage, Azure API Management, Azure Key Vault, Entra ID) AWS(SageMaker, Bedrock, Lambda, EC2, ECS, EKS, ECR, S3, OpenSearch, Glue, Athena, EMR, Textract, CloudWatch, CloudTrail, IAM, KMS) GCP(Vertex AI, Google Kubernetes Engine (GKE), BigQuery, Cloud Storage, Dataflow)
- **Data Engineering & Streaming:** Apache Spark, PySpark, Kafka, AWS Kinesis
- **Containers, DevOps & IaC:** Docker, Kubernetes, Helm, Terraform, Ansible, Git, GitHub Actions, Jenkins, AWS CodeBuild
- **Monitoring & Observability:** Prometheus, Grafana, Azure Monitor, ELK stack, openTelemetry, datadog, Dynatrace
- **Security:** OAuth2, OpenID Connect, AWS IAM, Azure Entra ID, API Gateway Security, Azure Key Vault, AWS KMS, Encryption, Private Endpoints, Zero Trust
- **Visualization & BI:** Matplotlib, Seaborn, Tableau, Power BI, Looker, QlikView
- **IDEs & Developer Productivity:** VS Code, Cursor AI, IntelliJ IDEA, PyCharm, Jupyter Notebook, JupyterLab, GitHub Copilot, Postman, Swagger/OpenAPI, Docker Desktop, Bash, PowerShell
- **Agile & Collaboration:** Jira, Confluence, Agile, Scrum

Professional Experience

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

- Designed, developed, and deployed **enterprise GenAI solutions** for supply-chain planning and forecasting using **Azure OpenAI Service, Python, LangChain, LangGraph, and FastAPI**, enabling AI-assisted decision support for hundreds of internal users.
- Implemented **Agent-to-Agent (A2A) workflows** using **LangGraph**, coordinating retrieval, reasoning, and validation agents for complex multi-step supply-chain decision scenarios.
- Designed and implemented **Retrieval-Augmented Generation (RAG) pipelines** using **Azure Cognitive Search (vector indexing), Azure OpenAI embeddings, Azure Blob Storage, and semantic chunking** to enable contextual Q&A over SOPs, planning documents, and operational data.
- Applied **LoRA and QLoRA parameter-efficient fine-tuning** on pretrained LLMs using **Hugging Face and PyTorch** to adapt models for domain-specific supply-chain terminology while minimizing compute and cost overhead.
- Developed **scalable backend AI microservices** using **Python, Go, FastAPI, Azure Functions, and Azure App Service**, exposing secure REST APIs consumed by internal web and analytics applications.
- Integrated **Hugging Face Transformers and PyTorch-based embedding models** with Azure OpenAI, improving semantic retrieval accuracy across large operational document repositories.
- Evaluated and implemented **vector database strategies** using **Pinecone and Weaviate** to optimize enterprise-scale semantic search and retrieval performance.
- Engineered **data ingestion and preprocessing workflows** using **Azure Data Factory, Azure Databricks (PySpark), Pandas, and Azure Blob Storage**, converting structured ERP data and unstructured PDFs into vectorized knowledge stores.
- Optimized LLM behavior through **advanced prompt engineering**, few-shot learning, temperature tuning, token control, and output validation, reducing hallucinations in production by **38%**.
- Integrated GenAI services with enterprise systems using **REST APIs, OAuth2, Azure Entra ID, Managed Identities, and Azure API Management**, enforcing role-based access control and security standards.
- Containerized and deployed GenAI workloads using **Docker, Azure Kubernetes Service (AKS), Azure Container Registry (ACR), Helm, Git, and GitHub Actions**, implementing CI/CD pipelines aligned with release governance.
- Implemented monitoring, observability, and production validation using **Prometheus, Grafana, unit and integration testing, prompt regression testing, and UAT**, while securing platforms with **Azure Key Vault, private endpoints, and encryption**, and documenting architectures in **Confluence**.

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

- Led the design and implementation of enterprise AI and GenAI solutions for banking operations using **Amazon SageMaker, Amazon Bedrock, Python, Hugging Face Transformers, LangChain, and LangGraph**, enabling intelligent automation across regulated UBS platforms.
- Designed **Agent-to-Agent (A2A) workflows** using **LangGraph**, coordinating policy-retrieval, compliance-reasoning, and response-validation agents for multi-step decision-making in regulated environments.
- Applied **LoRA and QLoRA fine-tuning** on pretrained LLMs to adapt models for financial and compliance-specific language, improving response precision while maintaining cost efficiency and governance.
- Built production-grade AI and GenAI services primarily in **Python**, with selective use of **Java and Scala** for high-throughput data processing and scalable model-serving pipelines.
- Designed and implemented **Retrieval-Augmented Generation (RAG)** pipelines using **Amazon OpenSearch (vector search), FAISS, Amazon S3, and embedding models**, enabling secure contextual search over policies and historical case data.
- Developed scalable backend AI microservices using **Python, FastAPI, Flask, Docker, and AWS Lambda**, serving millions of inference requests monthly with sub-second latency.

- Leveraged **Neo4j knowledge graphs** to model entity relationships and enhance context-aware retrieval within RAG-based compliance and policy intelligence systems.
- Supported backend platform services written in **Go** for API gateways, internal tooling, and observability components integrated with Python-based AI inference services.
- Engineered document intelligence pipelines using **Amazon Textract, NLP preprocessing, semantic chunking, and embedding generation**, reducing document processing time by **48%**.
- Optimized LLM behavior through **advanced prompt engineering**, few-shot learning, temperature tuning, token budgeting, response validation, and guardrail logic to ensure compliant and controlled outputs.
- Deployed and governed AI workloads using **Amazon EC2, ECS, EKS, ECR, Docker, and Terraform**, implementing CI/CD, change management, and release governance aligned with UBS standards.
- Implemented monitoring, logging, and compliance observability using **Amazon CloudWatch, AWS CloudTrail, OpenSearch Dashboards**, and centralized audit logging, ensuring traceability and regulatory readiness.
- Collaborated cross-functionally with data scientists, risk managers, compliance officers, and product owners using **Agile/Scrum methodologies, Jira, and Confluence**, aligning AI delivery with UBS regulatory and business objectives.
- Validated and governed AI systems through **unit testing, integration testing, model performance benchmarking, bias checks, human-in-the-loop review, and UAT cycles**, ensuring adherence to UBS risk and compliance policies.
- Secured and documented AI platforms by implementing **AWS Secrets Manager, AWS KMS encryption, private VPC networking**, and maintaining architecture diagrams, runbooks, and operational documentation in **Confluence**.

Landis +Gyr, Atlanta, GA ,USA | Fullstack AI Engineer | January 2024 – May 2024

- Delivered end-to-end **AI, ML, and GenAI solutions** for smart-energy analytics platforms by combining **LLM-powered insights, predictive modeling, and real-time IoT data processing** to support grid operations and energy consumption analysis.
- Developed autonomous agent workflows using **LangChain** and **CrewAI** to orchestrate data retrieval, anomaly analysis, and insight generation for smart-energy operations.
- Developed **LLM-enabled applications** using **Amazon Bedrock (Claude/Llama models), Hugging Face Transformers, Python, and LangChain**, enabling natural-language interaction with smart-meter data, outage reports, and operational documentation.
- Applied **pretrained LLMs** for energy analytics use cases, combining **RAG, agent workflows, and ML inference pipelines** without custom foundation-model training.
- Implemented **Retrieval-Augmented Generation (RAG) pipelines** using **Amazon OpenSearch (vector search), FAISS, Amazon S3, embedding models, and semantic chunking**, enabling accurate contextual querying of historical grid data and maintenance logs.
- Built **agent-style AI workflows** using **LangChain, prompt engineering, few-shot learning, response validation, and tool orchestration**, enabling autonomous data retrieval, summarization, and actionable insight generation.
- Designed and trained **machine learning and deep learning models** for **load forecasting, anomaly detection, and consumption trend analysis** using **Amazon SageMaker, XGBoost, TensorFlow, Pandas, NumPy,** and time-series feature engineering improving forecast accuracy by **14%** over baseline statistical models.
- Developed **backend AI and GenAI microservices** using **Python FastAPI, Java Spring Boot, REST APIs, and SQL/NoSQL databases**, exposing ML predictions and LLM outputs to internal dashboards and operational tools.
- Engineered **high-volume data ingestion pipelines** using **AWS IoT Core, Amazon Kinesis Data Streams, AWS Glue, Amazon S3, and Apache Spark on Amazon EMR**, supporting both real-time and batch ML workloads.
- Utilized **PyTorch models** and **Hugging Face libraries** to support forecasting, anomaly detection, and natural-language querying of operational and IoT datasets.
- Operationalized AI and GenAI models using **MLOps best practices**, including **SageMaker Pipelines, model versioning, experiment tracking, automated retraining, and environment promotion**.
- Deployed AI and ML services on **Kubernetes** across **AWS** environments, enabling scalable inference and resilient cloud-native model-serving architectures.
- Containerized and deployed AI services using **Docker, Amazon ECS, Amazon EKS, Amazon EC2, and Amazon ECR**, ensuring scalable, resilient, and low-latency inference in production environments.
- Implemented **CI/CD and MLOps automation** using **Git, GitHub Actions, AWS CodeBuild, and infrastructure-as-code**, enabling controlled and repeatable deployments across environments.
- Monitored and optimized AI systems using **Amazon CloudWatch, custom ML metrics, inference latency monitoring, and data-drift detection**, proactively identifying performance and data-quality issues.
- Documented and governed AI platforms by maintaining **model cards, architecture diagrams, deployment runbooks, and operational SOPs in Confluence**, supporting audits, maintainability, and cross-team onboarding.

UHG, Dallas, Texas, USA | Python Full Stack Engineer | September 2023 – December 2023

- Developed scalable **full-stack healthcare applications and backend services** using **Python, Flask, Django, Pandas, and NumPy**, supporting ingestion and processing of structured and unstructured clinical data.
- Built and integrated **RESTful APIs** to support healthcare workflows, enabling seamless data exchange between frontend applications and backend AI/analytics services.
- Designed and implemented **machine learning and deep learning models** including **CNNs for medical imaging, RNNs/LSTMs for sequential health data, GANs for synthetic data generation, and Transformer-based architectures** for NLP-driven use cases.
- Built and operationalized **end-to-end ML pipelines** using **Kubeflow on Google Kubernetes Engine (GKE)**, enabling reproducible training, deployment, and monitoring of predictive models.
- Leveraged **Google Vertex AI** for model experimentation, versioning, training, and scalable deployment of healthcare ML and AI workloads.
- Performed **time-series forecasting and trend analysis** using **ARIMA and Facebook Prophet** to predict patient utilization patterns, outcomes, and seasonal healthcare trends.

- Applied **Generative AI and NLP models** such as **GPT and BERT via Hugging Face**, enhancing patient communication workflows and extracting insights from clinical notes and feedback.
- Implemented **supervised machine learning algorithms** including **Random Forests, Support Vector Machines, and Neural Networks** to analyze high-dimensional healthcare datasets.
- Developed **computer vision pipelines** using **OpenCV, YOLO, and ImageNet pre-trained models**, supporting diagnostic image analysis and care prediction features.
- Utilized **GCP data services** including **BigQuery, Cloud Storage, and Dataflow** to store, process, and analyze large-scale healthcare datasets.
- Implemented **ML experiment tracking and model lifecycle management** using **MLflow**, ensuring reproducibility, traceability, and performance comparison across iterations.
- Improved **patient segmentation, retention, and treatment targeting** using **unsupervised learning techniques** such as **k-means clustering and PCA**, and visualized insights using **Matplotlib, Seaborn, Tableau, Power BI, Looker, and QlikView**.

Celanese, Irving, Texas, USA |Python Developer | May 2023 – August 2023

- Designed and developed **enterprise-grade microservices** using **Python, Flask, Django, and FastAPI**, supporting scalable backend systems and business workflows.
- Deployed **high-availability backend services** using **Azure Functions, Azure Kubernetes Service (AKS), and serverless architectures**, ensuring fault tolerance and horizontal scalability.
- Implemented **API gateway solutions** using **Azure API Management**, handling authentication, request routing, rate limiting, throttling, and load balancing across microservices.
- Built **event-driven architectures** using **Apache Kafka, Azure Event Hubs, and Azure Service Bus**, enabling real-time data streaming and asynchronous processing.
- Developed and maintained **data ingestion and ETL pipelines** integrating **Azure Data Factory**, SQL databases, NoSQL stores, external APIs, and IoT data sources.
- Engineered **CI/CD pipelines** using **Azure DevOps, GitHub Actions, Terraform, and Ansible**, automating application deployment and infrastructure provisioning.
- Implemented **asynchronous background processing** using **Celery and Azure Cache for Redis**, improving system throughput and API performance.
- Built **real-time communication services** using **Django Channels, WebSockets, Socket.IO, and Azure Cache for Redis**, enabling live notifications and messaging.
- Integrated and optimized **GraphQL APIs** within Azure-hosted services, improving backend query efficiency and reducing data over-fetching.
- Performed **database performance tuning and query optimization** across **Azure PostgreSQL, Azure MySQL, Cassandra, and Azure Cognitive Search / Elasticsearch**, applying indexing and schema optimization.
- Containerized applications using **Docker** and deployed them on **Azure Kubernetes Service (AKS)**, ensuring consistent environments and scalable deployments.
- Developed **AI-enabled backend services** on Azure, integrating **machine learning models, analytics components, and Azure AI services** into enterprise business intelligence applications.

Algocode, Pune, India | Python Developer | August 2020 – July 2022

- Developed responsive and user-friendly **web interfaces** using **HTML5, CSS3, JavaScript, and jQuery**, ensuring cross-browser compatibility and intuitive user experiences.
- Designed and implemented **full-stack web applications** using **Python and Django**, building frontend components and backend business logic for enterprise platforms.
- Enhanced **data exploration and filtering** using **PyQt**, enabling efficient navigation and analysis of financial transactions and statement data through rich UI components.
- Built and maintained **internal administrative tools** using **Django, uWSGI, and SQL**, supporting application management and **BDD-based workflows**.
- Integrated **Query UI components** with **Python and Django** to manage content lifecycle operations, including secure storage, retrieval, and deletion.
- Established **automated CI pipelines** using **Git, Jenkins, MySQL, Bash, and Python scripts**, enabling consistent builds, testing, and deployments.
- Deployed and supported applications in **Linux environments**, gaining hands-on experience with shell scripting, system monitoring, and command-line tools.
- Worked with **AWS services** including **EC2, S3, IAM, and RDS**, and supported orchestration and data pipelines using **AWS Glue, Step Functions, and AWS Data Pipeline**.
- Developed **Python automation scripts** to ingest, process, and analyze diverse data formats such as **structured feeds, XLS files, FIXML, and system logs**.
- Utilized **Apache Spark and Spark SQL** for data integration and transformation, contributing to scalable data-processing workflows.
- Applied **data analysis and early ML techniques** using **NumPy, SciPy, Matplotlib, H2O, and MLlib** for exploratory analytics and proof-of-concept use cases.
- Managed source code, sprint tasks, and collaboration using **Git, GitLab, and Jira**, following Agile practices and version-control best standards.

Education

University of North Texas, Denton, TX - Master's in information systems and technology | August 2022- May 2024

Certifications:

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

Work Authorization: Authorized to work in the U.S. (W2)