SAI THOTA

AI/Machine Learning Engineer

saithota5897@gmail.com| +1 (469) 573-2313 | LinkedIn| GitHub

databricks Generative AI Fundamentals

Microsoft CERTIFIED AZURE AI ENGINEER ASSOCIATE

aws certified Cloud Practitioner FOUNDATIONAL

PROFESSIONAL SUMMARY

- **Over 8 years** of experience delivering production-grade **AI/ML solutions** across Insurance, Cybersecurity, Healthcare, and Finance using Python, FastAPI, LangChain, AutoGen, and Docker.
- Specialized in designing **multi-agent AI** architectures with **RAG pipelines,** agent evaluation & monitoring frameworks, HITL mechanisms, and secure, cost-efficient private **LLaMA 3 /vLLM** stacks on AKS for scalable, compliant inference.
- Built secure AI pipelines for **multimodal threat detection** using **PyTorch**, OpenCV and **Weaviate VectorDB**, combined with **LangChain-powered RAG** pipelines and **self-hosted LLMs** for real-time semantic threat analysis.
- Proficient in **fine-tuning** and deploying **transformer-based and DL models** (MesoNet, ResNeXt-LSTM, WaveNet, GANs) with LoRa/QLoRA, PEFT, and adversarial training for robust detection in multilingual and low-resolution environments.
- Skilled in building large-scale data ingestion and processing pipelines using **AWS** (S3, Lambda, SageMaker, Glue), **Databricks** (PySpark, Delta Lake), **Kafka**, Airflow, and **Azure ML** with AKS for secure and compliant AI deployments.
- Developed **ML pipelines** with LightGBM, XGBoost, NLP (VADER, Logistic Regression), clustering, and recommendation systems to improve patient satisfaction and reduce operational costs by $120K/month.
- Strong **MLOps** expertise including Azure ML, MLflow, **Docker, Kubernetes,** Azure DevOps, GitHub Actions, **CI/CD** pipelines, secure FastAPI (**OAuth2/JWT)**, agent telemetry monitoring, and drift detection for audit-ready deployments.
- Delivered **secure, compliant AI/ML** solutions with strict adherence to HIPAA and GDPR regulations, incorporating PHI/PII-safe data handling, encrypted data transit, IAM-based access controls, audit logging, and human-in-the-loop mechanisms for risk mitigation.
- Applied advanced **prompt engineering** and **agent evaluation** frameworks to reduce LLM hallucinations by 17% and improve AI workflow reliability by 22%.
- Delivered **impactful AI solutions** enabling 38% reduction in manual claims review, 35% faster underwriting decisions, 20% faster fraud detection, and 30% overall operational efficiency improvements.

SKILLS

Category	Technologies
Languages & Scripting	Python, SQL, R, PySpark, Bash, PowerShell
ML & AI Frameworks	Scikit-learn, TensorFlow, PyTorch, XGBoost, LightGBM, pandas, NumPy, SciPy
Generative AI & LLMs	GPT-3.5, GPT-4, Claude, Amazon Titan, LLaMA 3, vLLM, BERT, RoBERTa, DeBERTa, T5, SentenceTransformers, QLoRA, LoRA, LangChain, LangGraph, AutoGen, Mistral, RAG, PEFT, Prompt Engineering, Agent Evaluation & Monitoring
NLP & Reasoning	Hugging Face Transformers, TF-IDF, VADER, Semantic Routing, In-context Learning (zero-shot, one-shot)
Time Series & Forecasting	Random Forest Regressor (RFR), Seasonal Decomposition, Graph Neural Networks (GNN), Prophet, ARIMA, SARIMA, LSTM, BiLSTM
Cloud Platforms	AWS (S3, Lambda, EC2, SageMaker, Step Functions, Glue, API Gateway, CloudWatch, IAM), Azure ML, Azure Kubernetes Service (AKS), Azure DevOps, Azure Monitor, OpenTelemetry
MLOps and DevOps Tools	MLflow, Docker, Kubernetes, FastAPI (OAuth2/JWT Auth), Flask, DVC, Terraform, GitHub Actions, GitLab, Jenkins, Bitbucket, CI/CD Pipelines, Drift Detection
Data Engineering & ETL	PySpark, Apache Kafka, Airflow, SQLAlchemy, AWS Glue, Lambda, Redis, MongoDB, Delta Lake, ChromaDB, FAISS, Qdrant, Pinecone, Weaviate, SQL Server, Cosmos DB, SSRS
Data Acquisition & Web Scraping	BeautifulSoup, urllib, requests, Selenium, APIs (Google Maps, USBR, USGS, SNOTEL)
Visualization & BI	Power BI, Tableau, Plotly, Matplotlib, Seaborn, Streamlit, EXCEL (Power Query, DAX)
Tools & Collaboration	Jupyter, VS Code, PyCharm, Git, Confluence, JIRA, Agile/Scrum
Domain Expertise	Healthcare Analytics, Cybersecurity & Threat Detection, Insurance Claims & Underwriting, Regulatory Knowledge Bases, Water Resource Forecasting, Risk & Fraud Analytics

PROFESSIONAL EXPERIENCE

Client: Allied World Assurance Company (AWAC) | New York, NY **October 2024 – Present**
Role: AI/ML Engineer

Description: Built an agentic AI architecture combining multi-agent orchestration and RAG pipelines to automate claims processing, fraud detection, and underwriting intelligence. Implemented a hybrid approach—prototyping with OpenAI APIs and migrating to a private LLaMA 3 stack on Azure for secure, cost-efficient production deployments.

Key Contributions:

- Designed and implemented claims automation workflows using **LangChain** agents and **LangFlow** for document ingestion, fraud scoring, and policy lookups, reducing manual review effort by **38%** and accelerating claims triage.
- Developed an underwriting risk intelligence assistant using **RAG pipelines** with **ChromaDB** and regulatory knowledge bases, reducing decision turnaround by **35%** and improving quote accuracy.
- Established robust **agent evaluation & monitoring** pipelines inspired by **Arize/LLM-as-Judge** techniques, enabling traceability of agent reasoning, prompt versioning, and boosting agent reliability scores by **22%.**
- Integrated telemetry and monitoring using **Azure Monitor** and **OpenTelemetry** for real-time observability and drift detection, supporting proactive model management and audit readiness.
- Implemented human-in-the-loop (**HITL**) mechanisms and fallback flows to mitigate risk from hallucinations or uncertain outputs, ensuring safe, compliant decision-making.
- Migrated prototypes from **OpenAI function calling** to a private **LLaMA 3** inference stack deployed on Azure Kubernetes Service (**AKS**) using **vLLM**, incorporating model optimization techniques like **quantization** and **LoRA** to reduce inference costs by **23%** while enforcing PHI/PII-safe handling.
- Delivered secure **CI/CD** pipelines with **Azure ML, Azure DevOps, MLflow, and FastAPI** microservices, integrating **OAuth2/JWT** authentication, feature stores, and automated drift detection for audit-ready deployments.
- Collaborated closely with claims operations, underwriting SMEs, and compliance teams through agile sprints, iteratively refining POCs, addressing edge cases, and driving production readiness.

Technical Stack: LangChain, LangGraph, LangFlow, AutoGen, Hugging Face Transformers (DeBERTa, T5), LLaMA 3, vLLM, OpenAI Function Calling, ChromaDB, Azure ML, Azure Kubernetes Service, Azure DevOps, MLflow v2.14, Docker, SQL Server, Cosmos DB (4.2.0), FastAPI, Prompt Engineering, Agent Evaluation, RAG, MLOps, Secure AI Deployment.

Client: McAfee | Frisco, TX **May 2023 – October 2024**
Role: Data Scientist – LLMs & Generative AI

Description: Developed a secure, enterprise-scale AI pipeline for detecting deepfake media, zero-day malware URLs, and emerging cyber threats during the 2024 U.S. election cycle. Leveraged multimodal ML models, GenAI-powered RAG pipelines, and secure data lakes for real-time threat detection and narrative summarization, with all data and inference fully contained within McAfee's secure cloud.

Key Contributions:

- Designed and implemented secure data ingestion pipelines using **AWS VPC Peering, Kafka streaming, Delta Lake**, and **Databricks (PySpark, MLflow)** to process petabyte-scale telemetry under strict **AWS RBAC** controls.
- Preprocessed **multimodal** (video/audio) data with **PyTorch** and **OpenCV**, performing frame stabilization, voice isolation, and **feature extraction** (MFCC, FFT, DCT), with anonymized embeddings stored in a self-hosted **Weaviate** VectorDB.
- Fine-tuned and deployed **MesoNet, ResNeXt-LSTM, WaveNet, and CNN-BiLSTM models** for forgery detection, video temporal consistency, voice cloning, and speech anomaly detection, improving robustness with **GAN-generated** adversarial samples.
- Integrated self-hosted **Mistral LLM** inference endpoints via **secure APIs and LangChain** powered **RAG** pipelines to validate content and enable semantic threat analysis.
- Reduced **LLM hallucinations up to 17%** with **LangSmith prompt optimization**, improving the accuracy and consistency of multilingual threat narratives and achieving **87%+ detection accuracy** with an **18% reduction in false negatives**.
- Automated alerting and incident response workflows through **Databricks jobs, AWS SageMaker** inference endpoints, and Lambda orchestration, maintaining real-time threat mitigation within McAfee's compliance perimeter.
- Ensured **GDPR and HIPAA** compliant handling of PHI/PII data via IAM-based access control, encrypted data transit, audit logging, and FastAPI microservices secured with **OAuth2/JWT**.

Technical Stack: Python, PyTorch, FastAPI, Deep Learning Models, Attention, GAN, Databricks (PySpark, MLflow), Weaviate (self-hosted), Mistral (self-hosted LLM), LangChain, AWS SageMaker, AWS Lambda, Kafka, Delta Lake, GenAI, SHAP, LIME, OAuth2/JWT, GenAI, RAG, Multimodal Threat Detection, RAG, Secure AI Inference, DFDC, Hallucination Detection.

Client: GE HealthCare | Chicago, IL, USA **July 2019 - August 2022**
Role: Senior Data Engineer

Description: Redesigned data infrastructure and built production-ready ML/analytics pipelines to modernize GE HealthCare's mobile event logistics and patient feedback systems. Led API data ingestion, feature engineering, model deployment, and microservice integration to support scalable routing optimization and patient experience analytics.

Key Contributions:

- Engineered data pipelines to replace static aerial-distance calculations with dynamic route estimation using **Google Maps APIs**, reduced data latency and enabled **$120K/month cost savings** in planning for mobile healthcare events.
- Designed and trained **LightGBM** models (benchmarked with **XGBoost**) using 7 months of route data to predict real-world travel distances with over **95% accuracy**, reducing third-party API reliance and improving scheduling efficiency.
- Built a scalable patient feedback analytics pipeline using **spaCy, NLTK, VADER**, and **Logistic Regression** for sentiment analysis, applied **K-Means** clustering, **Apriori rule mining**, and collaborative filtering on provider-patient interaction data, improving satisfaction scores by **0.3 in 6 months.**
- Developed robust **ETL pipelines** and **batch jobs** in Python, SSRS, and SQL to automate data ingestion, transformation, and reporting across patient feedback, logistics, and revenue systems.
- Deployed analytics microservices and model endpoints using **FastAPI, Docker, and AWS** (Lambda, EC2, S3), containerized pipelines integrated into scheduling systems via **REST APIs**.
- Collaborated with stakeholders, logistics, revenue, and onsite data collection teams to translate business logic into scalable pipelines, following **Agile methodology** and **HIPAA-compliant** development standards.

Technical Stack: Python (pandas, NumPy, scikit-learn, spaCy, NLTK, openpyxl, SQLAlchemy), SQL Server, SSRS, LightGBM, XGBoost, VADER, Logistic Regression, K-Means, Apriori, Collaborative Filtering, AWS (Lambda, EC2, S3), Docker, FastAPI, Git, Jenkins, Excel, REST APIs, Data Pipelines, ETL, Microservices, Agile, HIPAA Compliance.

Client: N-iX | Bengaluru, Karnataka, India **April 2017 - June 2019**
Role: Python Developer

Description: Built an IoT-enabled agriculture analytics system by combining field surveys with LoRa-based sensor telemetry for real-time monitoring. Applied Python-driven preprocessing, EDA, and modeling to reveal socioeconomic trends and digital infrastructure gaps. Delivered insights via Excel dashboards and interactive web apps, enhancing transparency and rural planning efficiency.

Key Contributions:

- Designed and executed field surveys in rural Indian villages to evaluate the impact of **IoT in agriculture**, collected and stored 1,000+ farmers records in **MS SQL Server** using custom schemas, tables, and stored procedures.
- Built Excel **dashboards** and pivot tables to report key agricultural, socioeconomic, and other insights to stakeholders.
- Preprocessed data in **Python** using pandas, NumPy, and scikit-learn; applied **StandardScaler**, outlier handling, and conducted exploratory data analysis (**EDA**) to find rural infrastructure and digital gaps.
- Developed a **Flask-based web app** integrated with LoRa modules for real-time sensor data visualization and transmission over 0.5–1.9 miles in varied terrain mainly in agricultural fields.
- Modeled relational data using **SQLAlchemy**, improving query efficiency and ensuring data integrity across the pipeline.

Technical Stack: Questionnaire Design, Survey Execution, Data Collection & Processing, Python (pandas, NumPy, scikit-learn, matplotlib, Flask), SQL (SQL Server, SQLAlchemy, Window Functions), Excel (Power Query, Pivot Tables, XLOOKUP), Jupyter Notebook, Anaconda, EDA, GitHub, Stakeholder Communication, IoT, LoRa.

CERTIFICATIONS(Link)

- Databricks Generative AI Fundamentals
- Microsoft Azure AI Engineer Associate
- AWS Cloud Practitioner

EDUCATION

- **Master of Science in Computer Science** and **Data Science Graduate Certificate** at **Utah State University**, Logan, Utah