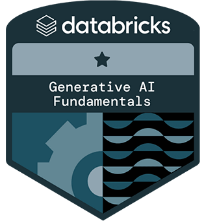
**A blue and white logo

AI-generated content may be incorrect.SAI THOTA**

**Senior AI/Machine Learning Engineer**

[**tsai58997@gmail.com**](mailto:tsai58997@gmail.com|)**|** +1 (469) 573-2313 | [**LinkedIn**](https://www.linkedin.com/in/tsai58997/)| [**GitHub**](https://github.com/TSai58997)

**AI/ML | multi-Agent AI | RAG | Generative AI | LangChain | LangFlow | vLLM | OpenAI | PyTorch |TensorFlow | NLP | Computer Vision | Time Series | Prompt Engineering | MLOps (Azure, AWS, CI/CD) | Vector DBs | Secure AI Deployment | Docker | Kubernetes | Databricks | Statistics**

**PROFESSIONAL SUMMARY**

* **Over 8 years** of experience delivering production-grade **AI/ML solutions** across Insurance, Cybersecurity, Healthcare, Environmental, and Manufacturing domains using Python, LangChain, Docker, and Databricks.
* Spearheaded design and deployment of **multi-agent AI systems** using **OpenAI GPT-4** and private **LLaMA 3 LLM** stacks with **LoRA/QLoRA** optimization, enabling scalable and efficient generative AI workflows.
* Leveraged **RAG pipelines**, prompt engineering, **agent** evaluation & monitoring, and **HITL** mechanisms with **ChromaDB** to ensure reliability, accuracy, and compliance in AI decision-making.
* Engineered secure, **multimodal** AI pipelines for **threat detection** using **PyTorch**, OpenCV and **Weaviate**, integrated with **LangChain-powered RAG** pipelines and **self-hosted LLMs** for real-time semantic analysis.
* Proficient in **fine-tuning** and deploying **transformer-based and deep learning models** (ResNeXt-LSTM, WaveNet, GANs), and adversarial training for robust detection in multilingual and low-resolution environments.
* Delivered **impactful AI solutions** that accelerated claims processing, underwriting decisions, fraud detection, and patient satisfaction while reducing operational costs and manual review efforts.
* Applied ensemble learning, **time-series forecasting, and graph-based models** (LSTM, GNN, SARIMA) to optimize predictions and improve decision-making in high-volume, dynamic data environments.
* Led large-scale **data engineering** programs, including ingestion, streaming, batch processing, ETL/ELT, and **warehousing** using **AWS** (S3, Lambda, SageMaker, Glue), **Databricks** (PySpark, Delta Lake), **Kafka**, and Airflow.
* Built enterprise-grade **MLOps** **workflows** using **Azure ML**, **AKS**, **MLflow**, **Docker, Kubernetes**, FastAPI, and CI/CD automation, integrating telemetry, drift detection, and PHI/PII-safe governance for audit-ready deployments.
* Directed **end-to-end data analytics and transformation** workflows, turning raw datasets into actionable insights using SQL, Python, and visualization tools, and dashboards for reporting and to improve operational efficiency.
* **Strong foundation** in statistical modeling, probability, and **mathematical** optimization to boost AI/ML model performance and reliability.

**SKILLS**

|  |  |
| --- | --- |
| **Category** | **Technologies** |
| **Languages & Scripting** | Python, SQL, R, PySpark, Bash, PowerShell |
| **ML & AI Frameworks** | Scikit‑learn, TensorFlow, PyTorch, XGBoost, LightGBM, pandas, NumPy, SciPy, SHAP, LIME |
| **Frameworks & Orchestration (Gen AI & LLMs)** | LangChain, LangGraph, LangFlow, LangSmith, AutoGen, OpenAI, RAG, PEFT, Prompt Engineering, Agent Evaluation & Monitoring |
| **Models & Fine-tuning**  **(Gen AI & LLMs)** | GPT‑3.5, GPT‑4, Claude, Amazon Titan, LLaMA 3, vLLM, Mistral, BERT, DeBERTa, SentenceTransformers, LoRA/QLoRA |
| **NLP & Reasoning** | TF‑IDF, VADER, Semantic Routing, In‑context Learning (zero‑shot, one‑shot) |
| **Vector Databases** | ChromaDB, FAISS, Qdrant, Pinecone, Weaviate |
| **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, Flask, DVC, Terraform, GitHub Actions, GitLab, Jenkins, Bitbucket, CI/CD, Drift Detection |
| **Data Engineering & ETL** | Databricks, PySpark, Apache Kafka, Airflow, SQLAlchemy, AWS (Glue, Lambda), Redis, MongoDB, Delta Lake, SQL Server, Cosmos DB, SSRS, SSIS, SSAS |
| **Data Acquisition &**  **Web Scraping** | BeautifulSoup, urllib, requests, Selenium, APIs (Google Maps, USBR, USGS, SNOTEL) |
| **Visualization & BI** | Power BI, Tableau, Plotly, Matplotlib, Seaborn, Streamlit, MS Excel (Power Query, DAX) |
| **Tools & Collaboration** | Jupyter, VS Code, PyCharm, Git, Confluence, JIRA, Agile/Scrum |
| **Domain Expertise** | Healthcare Analytics, Cybersecurity & Threat Detection, Regulatory Knowledge Bases, Insurance Claims & Underwriting, Water Resource Forecasting, Risk & Fraud Analytics |

**PROFESSIONAL EXPERIENCE**

**Client: Allied World Assurance Company (AWAC) | New York, NY October 2024 – Present**

**Role: Senior AI/ML Engineer**

**Description**: Built secure, agentic AI systems to automate claims processing, fraud detection, and underwriting risk assessment. Prototyped agentic AI workflows using OpenAI APIs for rapid development. Migrated production deployments to a private LLaMA 3 stack on Azure, ensuring secure, scalable, and cost-efficient inference. Leveraged RAG pipelines, and human-in-the-loop mechanisms to improve decision accuracy and reduce manual review.

**Key Contributions:**

* Designed and deployed **OpenAI GPT-4–powered conversational agents** and **private LLaMA 3 stacks** with **LoRA/QLoRA optimization** for policy lookups, fraud detection, and claims triage, reducing manual processing time by **38%.**
* Developed an **underwriting risk intelligence assistant** using **DeBERTa** **embeddings** in LangChain RAG pipelines with **ChromaDB** and regulatory knowledge bases, improving decision turnaround by **35%** and quote accuracy.
* Built **multi-agent AI systems** leveraging **RAG pipelines, LangChain, and LangFlow** orchestration, **prompt engineering**, and agent evaluation & monitoring to ensure reliable, real-time generative AI workflows.
* Introduced human-in-the-loop (**HITL**) mechanisms and fallback flows to mitigate risk from hallucinations or high-uncertain outputs, implemented safeguards for compliant decision-making.
* Migrated prototypes to **private LLaMA 3 inference** on AKS using vLLM, enabling secure, scalable ChatGPT-style deployment and reducing inference costs by 23% via quantization and LoRA/QLoRA optimizations.
* Implemented **LangFlow orchestration** to streamline multi-turn dialogs, document ingestion, and workflow automation.
* Established **agent evaluation & monitoring** pipelines inspired by **Arize/LLM-as-Judge** techniques, boosting reliability scores and ensuring prompt traceability**.**
* Engineered **secure** **CI/CD** **pipelines** with Azure ML, Azure DevOps, Docker, Kubernetes, and FastAPI microservices, feature stores, and automated drift detection.
* Defined observability and telemetry requirements and collaborated with deployment teams to integrate **Azure Monitor and OpenTelemetry** for real-time monitoring and drift detection

**Technical Stack:** LangChain, LangGraph, LangFlow, AutoGen, Hugging Face Transformers (DeBERTa), LLaMA 3, vLLM, OpenAI, GPT-4, LoRA/QLoRA, ChromaDB, HITL, Azure ML, AKS, Azure DevOps, MLflow, Docker, Kubernetes, SQL Server, Cosmos DB, Prompt Engineering, Agent Evaluation & Monitoring, RAG, MLOps, Agile, and Secure AI Deployment.

**Client: McAfee | Frisco, TX May 2023 – October 2024**

**Role: Data Scientist – LLMs & Generative AI**

**Description**: Developed a secure AI-driven threat detection platform to identify deepfake media, voice cloning, and emerging cyber threats in real time. Fine-tuned multimodal AI models, and integrated private LLMs for multilingual threat analysis, achieving high detection accuracy, reduced false negatives, and rapid incident response.

**Key Contributions:**

* **Integrated self-hosted Mistral LLM endpoints via LangChain-powered RAG pipelines** with **Weaviate** VectorDB for real-time semantic threat analysis of multimodal data (video/audio).
* **Reduced LLM hallucinations by 17%** using LangSmith prompt optimization, improving multilingual threat narrative accuracy and achieving 87%+ detection accuracy with an 18% reduction in false negatives.
* **Fine-tuned and deployed transformer-based and deep learning models** (ResNeXt-LSTM, WaveNet, CNN-BiLSTM) for forgery detection, video temporal consistency, voice cloning, and speech anomaly detection.
* Enhanced model robustness and generalization with **GAN-generated** adversarial samples across multilingual and low-resolution datasets.
* Preprocessed **multimodal** (video/audio) data with **PyTorch** and **OpenCV**, performing frame stabilization, voice isolation, and **feature extraction** (MFCC, FFT, DCT).
* Designed secure data ingestion pipelines using **AWS VPC Peering, Kafka streaming, Delta Lake**, and **Databricks** (PySpark, notebooks, jobs, MLflow experiments) to process petabyte-scale telemetry under strict **AWS RBAC** controls.
* Automated alerting and incident response workflows through **Databricks jobs, AWS Lambda,** and secure 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.**
* Partnered with claims operations, underwriting SMEs, and compliance teams through **agile sprints,** iteratively refining POCs, addressing edge cases, and delivering **production-ready data engineering** and AI/ML pipelines.

**Technical Stack:** Python, PyTorch, LangSmith, LangChain, Deep Learning Models, Attention, GAN, Databricks, PySpark, MLflow, Weaviate (self-hosted), Mistral (self-hosted LLM), OpenCV, AWS Lambda, Kafka, Delta Lake, OAuth2/JWT, GenAI, RAG, Multimodal Threat Detection, Secure AI Inference, DFDC, Hallucination Detection, GDPR, HIPAA, PHI, PII.

**Client: GE HealthCare | Chicago, IL, USA July 2019 - August 2022**

**Role: Senior Data Engineer**

**Description**: Delivered scalable data pipelines and predictive models to optimize healthcare event logistics, reducing costs and improving patient satisfaction. Built HIPAA-compliant analytics services for scheduling, feedback analysis, and operational reporting, integrating them seamlessly into core systems.

**Key Contributions:**

* Crafted 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.**
* 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.
* 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.
* Supported model interpretability using **SHAP/LIME** for LightGBM/XGBoost predictions in patient feedback and logistics pipelines.
* 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 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), FastAPI, Git, Jenkins, Excel, REST APIs, Data Pipelines, ETL, Microservices, Agile, HIPAA Compliance, SHAP, LIME, PHI, PII.

**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 to stakeholders, enabling data-informed strategies for improving agricultural productivity and rural connectivity.

**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.
* Worked with **device manufacturing teams** to define sensor requirements, test prototypes, and validate signal reliability under varying environmental and terrain conditions.
* Developed calibration scripts and automated quality checks for LoRa-based sensors before deployment.
* 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 uncover rural infrastructure 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, IoT Sensor Manufacturing Collaboration, Data Collection & Processing, Python (pandas, NumPy, scikit-learn, matplotlib, Flask), SQL (SQL Server, SQLAlchemy, Window Functions), Excel (Power Query, Pivot Tables, XLOOKUP), Jupyter Notebook, EDA, GitHub, Stakeholder Communication, LoRa IoT Devices.Top of Form

**CERTIFICATIONS(**[**Link**](https://www.linkedin.com/in/saithota97/details/certifications/)**)**

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

**EDUCATION**

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