

SAI VINUTHNA TATA

Sr. Gen AI/ML Engineer

Email: saivinuthnat@gmail.com | Contact: +1(469) 759-9088 | [LinkedIn](#)

PROFESSIONAL SUMMARY:

- Senior AI/ML Engineer specializing in **Generative AI, Agentic AI, Large Language Models (LLMs)**, and **Machine Learning** with 10+ years of experience designing and delivering scalable **AI, Data Science, and Cloud solutions** across Healthcare, Financial Services, Government, and Retail domains using **Microsoft Azure, AWS, and Google Cloud Platform (GCP)**.
- Extensive expertise in developing enterprise production-grade **Gen AI** applications using **OpenAI GPT-4o, Azure OpenAI, Amazon Bedrock, LangChain, LangGraph, LlamaIndex, Model Context Protocol (MCP)**, and **AI Agents** for intelligent automation, enterprise knowledge management, document intelligence and conversational AI.
- Strong experience architecting **Retrieval-Augmented Generation (RAG)** solutions using **Hybrid Search, Semantic Search, OpenAI Embeddings, Sentence Transformers, Pinecone, FAISS, Azure AI Search**, and document intelligence frameworks to deliver accurate, context-aware, and explainable AI solutions.
- Proven expertise in building end-to-end **Machine Learning and AI platforms**, including data ingestion, distributed processing, feature engineering, model development, evaluation, deployment, monitoring, and optimization using **Python, Apache Spark, PySpark, Databricks, Scikit-learn, TensorFlow, PyTorch, and MLflow**.
- Hands-on experience developing secure cloud-native **AI microservices** and scalable APIs using **Python, FastAPI, Docker, and Kubernetes**, enabling scalable production AI applications and intelligent business workflows across cloud environments.
- Experienced in implementing enterprise **MLOps and LLMOps** practices using **Azure Machine Learning, MLflow, Docker, Kubernetes, Azure DevOps, GitHub Actions, and Jenkins**, with expertise in automated CI/CD, model versioning, prompt evaluation, observability, and continuous performance monitoring.
- Strong background in Machine Learning, **Deep Learning, Natural Language Processing (NLP), Predictive Analytics, Recommendation Systems, Time-Series Forecasting, Fraud Detection, Anomaly Detection, Healthcare Analytics, and Explainable AI (SHAP)** delivering intelligent, data-driven business solutions.
- Experienced designing scalable **Data Engineering** platforms using **Apache Spark, PySpark, Apache Kafka, Azure Data Factory, Databricks, Delta Lake, Snowflake, SQL Server**, and scalable ETL/ELT pipelines supporting large-scale AI and analytics workloads.
- Proficient in building executive dashboards and analytics solutions using **Power BI, Tableau, SQL, and Python**, transforming complex business data into actionable business insights through interactive visualizations and KPI reporting.
- Collaborative engineering professional with extensive experience working in **Agile Scrum** environments, partnering with **Product Owners, Solution Architects, Data Engineers, Software Developers, Clinical Experts, Business Stakeholders**, and executive leadership to deliver secure, scalable, and production-ready enterprise AI solutions.

TECH STACK:

Programming & Web Technologies	Python, Java, SQL, JavaScript, TypeScript, HTML5, CSS3, Shell Scripting, FastAPI, Flask, Django, Spring Boot, React.js, jQuery, Bootstrap, AJAX, REST APIs, OpenAPI
Generative AI & Agentic AI	OpenAI GPT-4o, Azure OpenAI, Anthropic Claude, Google Gemini, Amazon Bedrock, LangChain, LangGraph, LlamaIndex, LLM, Model Context Protocol (MCP), MCP Tools Integration, Agentic AI, Multi-Agent Systems, AI Agents, AI Copilots, Prompt Engineering, Function Calling, Structured Outputs
RAG, NLP & Document Intelligence	Retrieval-Augmented Generation (RAG), Hybrid Search, Semantic Search, OpenAI Embeddings, Sentence Transformers, Pinecone, FAISS, ChromaDB, Hugging Face Transformers, BERT, spaCy, NLTK, TF-IDF, OCR, Azure AI Search, Azure AI Document Intelligence, HL7/FHIR
Machine Learning & Deep Learning	Scikit-learn, TensorFlow, PyTorch, XGBoost, Random Forest, Gradient Boosting, Decision Trees, Logistic Regression, Support Vector Machines (SVM), K-Means Clustering, Principal Component Analysis (PCA), Regression, Linear Regression, Classification, Clustering, Recommendation Systems,

	Time-Series Forecasting, ARIMA, Holt-Winters, Feature Engineering, Hyperparameter Tuning, Model Evaluation, Explainable AI (XAI)
MLOps, LLMops & DevOps	MLflow, Azure Machine Learning, Amazon SageMaker, Docker, Kubernetes, Azure Kubernetes Service (AKS), Azure DevOps, GitHub Actions, Jenkins, CI/CD, Model Versioning, Model Monitoring, Prompt Evaluation, RAGAS, Langfuse, Git, GitHub, Linux
Data Engineering & Big Data	Apache Spark, PySpark, Spark MLlib, Databricks, Azure Databricks, Apache Airflow, Azure Data Factory, Apache Kafka, Spark Streaming, Delta Lake, BigQuery, ETL/ELT Pipelines, Data Warehousing
Cloud Platforms & Services	Microsoft Azure, Azure OpenAI, Azure AI Search, Azure Machine Learning, Azure Blob Storage, Azure Data Lake Storage, Azure Functions, Azure API Management, Azure Key Vault, Azure Monitor, Application Insights, Amazon SageMaker, AWS Glue, AWS Lambda, Amazon S3, Amazon Bedrock, Amazon Redshift, Google Cloud Platform (GCP), Google AI Platform, BigQuery, Cloud Storage, Cloud Dataproc
Databases & Analytics	SQL, SQL Server, PostgreSQL, MySQL, Oracle, Snowflake, MongoDB, Azure SQL Database, Azure Cosmos DB, Power BI, Tableau, Matplotlib, Microsoft Excel
Security, Monitoring & Methodologies	OAuth2, JWT, RBAC, HIPAA, PII Masking, Responsible AI, Azure AI Content Safety, SHAP, OpenTelemetry, Prometheus, Grafana, Postman, Agile Scrum, SDLC, Microservices Architecture, Design Patterns

WORK EXPERIENCE:

Client: Quest Diagnostics, Dallas, Texas

Senior AI/ML Engineer (Generative AI & Agentic AI)

Aug 2023 – Present

- Architected scalable **Gen AI** platforms on **Microsoft Azure Cloud Platform** using **LLMs, Agentic AI, RAG**, and workflow automation to modernize clinical decision support laboratory knowledge management, and enterprise healthcare intelligence.
- Designed scalable **multi-agent AI** solutions using **LangGraph, LangChain**, and **LlamaIndex** on **Azure Cloud Platform** for autonomous reasoning, memory management, enterprise healthcare workflow orchestration and intelligent task execution.
- Built production-ready **AI Copilots** using **GPT-4o, Azure OpenAI**, and **Azure AI Studio** for clinical document understanding, provider assistance, intelligent search, contextual healthcare question answering, and clinical decision support.
- Developed enterprise **RAG** frameworks integrating **EHRs**, laboratory reports, clinical documentation, and medical guidelines using **Azure AI Search**, hybrid retrieval, reranking, and semantic search for accurate knowledge retrieval.
- Engineered document intelligence pipelines using **Azure AI Document Intelligence, OCR**, semantic chunking, **Azure Blob Storage**, and metadata extraction to process physician notes, **HL7/FHIR** resources, and insurance documents.
- Designed scalable embedding and vector search pipelines using **OpenAI Embeddings, Azure AI Search, Pinecone**, and **FAISS** for high-performance semantic retrieval, contextual knowledge discovery and low-latency vector search.
- Implemented advanced **Prompt Engineering** strategies including few-shot learning, dynamic prompting, structured outputs, function calling, and response grounding to improve **LLM** accuracy, consistency, explainability, and reliability.
- Developed secure AI orchestration services using **Python, FastAPI, REST APIs, Azure Functions**, and microservices on **Azure Cloud Platform**, enabling seamless enterprise AI integration with healthcare applications and distributed AI services.
- Built distributed data pipelines using **PySpark, Azure Databricks, Azure Data Factory, SQL, Delta Lake**, and **Azure Data Lake Storage** supporting AI model training, analytics, feature engineering, and large-scale data processing.
- Developed predictive **Machine Learning** models using **Scikit-learn, XGBoost, TensorFlow, PyTorch**, and **Azure Machine Learning** for patient risk prediction, anomaly detection, operational forecasting, and healthcare analytics.
- Designed enterprise **LLMOps** and **MLOps** pipelines using **MLflow, Azure Machine Learning, Docker, AKS, Azure DevOps, GitHub Actions**, and **Jenkins** for automated deployment and lifecycle management.
- Implemented comprehensive **AI evaluation** using **RAGAS, Langfuse, LLM-as-a-Judge**, retrieval evaluation, hallucination detection, and human feedback to continuously improve production **LLM** performance, reliability, and response accuracy.
- Integrated **Azure OpenAI, Azure AI Search, Azure Machine Learning, Azure Key Vault, Azure API Management, AKS, Azure Functions**, and **Amazon Bedrock** with secure enterprise authentication and RBAC controls.

- Implemented **Model Context Protocol (MCP)** architectures connecting **LLMs** with enterprise APIs, healthcare systems, external knowledge repositories, and internal business applications across **Azure Cloud Platform** services using secure integrations.
- Designed enterprise observability solutions using **Azure Monitor**, **Application Insights**, **OpenTelemetry**, **Prometheus**, and **Grafana** to monitor model performance, infrastructure health, retrieval quality, API latency, and token utilization.
- Collaborated with physicians, laboratory specialists, product owners, enterprise architects, and security teams to deliver scalable **Generative AI** solutions while ensuring **HIPAA** compliance and healthcare governance standards.
- Implemented **Responsible AI** practices including guardrails, **PII masking**, prompt injection protection, audit logging, explainability, governance frameworks, and **Azure AI Content Safety** for secure enterprise healthcare AI deployments.

Environment: Python, SQL, FastAPI, OpenAI GPT-4o, Azure OpenAI, Anthropic Claude, Google Gemini, Amazon Bedrock, LangChain, LangGraph, LlamaIndex, Model Context Protocol (MCP), Agentic AI, Multi-Agent Systems, Retrieval-Augmented Generation (RAG), Prompt Engineering, Function Calling, Structured Outputs, OpenAI Embeddings, Azure AI Search, Pinecone, FAISS, Azure AI Document Intelligence, MLflow, Azure Machine Learning, Azure Databricks, Apache Spark, PySpark, Azure Data Factory, Delta Lake, Azure Blob Storage, Scikit-learn, XGBoost, TensorFlow, PyTorch, OCR, Docker, Kubernetes, Azure Kubernetes Service (AKS), Azure DevOps, GitHub Actions, Jenkins, Azure Monitor, Application Insights, OpenTelemetry, Prometheus, Grafana, Power BI, HL7/FHIR, HIPAA, REST APIs, Agile Scrum.

Client: BHG Financial, Syracuse, NY

Applied AI / Machine Learning Engineer

Feb 2021– July 2023

- Architected scalable **Applied AI** and **Machine Learning** solutions using **AWS Cloud**, **Python**, and **Scikit-learn** for credit risk assessment, underwriting analytics, fraud detection, and lending decision automation.
- Engineered end-to-end **Machine Learning** pipelines using **Apache Spark**, **PySpark**, **Databricks**, **Amazon SageMaker**, and **MLflow** for automated data ingestion, feature engineering, model training, deployment, monitoring, and lifecycle management.
- Developed predictive **Machine Learning** models using **XGBoost**, **Random Forest**, **Gradient Boosting**, and **Logistic Regression** for classification, regression, default prediction, customer risk profiling, and creditworthiness assessment.
- Designed intelligent **recommendation engines** using **collaborative filtering**, **clustering**, and **behavioral analytics** to deliver personalized **lending offers**, **refinancing recommendations**, **customer segmentation**, and **cross-selling opportunities**.
- Implemented enterprise **Natural Language Processing (NLP)** solutions using **spaCy**, **NLTK**, **Hugging Face Transformers**, and **AWS Comprehend** to analyze **loan documents**, **underwriting reports**, and **customer communications**.
- Built intelligent document processing pipelines using **AWS Textract**, **OCR**, metadata extraction, and validation rules to automate loan applications, identity verification, bank statements, tax documents, and regulatory compliance workflows.
- Implemented **Explainable AI (XAI)** frameworks using **SHAP**, feature importance analysis, statistical validation, and model interpretability techniques to improve regulatory compliance, model transparency, and stakeholder confidence.
- Developed scalable **AI microservices** using Python, **FastAPI**, **Flask**, **Docker**, and **Amazon API Gateway** for secure real-time model inference, API integration, and enterprise lending applications across distributed environments.
- Engineered distributed data pipelines using **Apache Spark**, **AWS Glue**, **Amazon S3**, **Snowflake**, and **SQL Server** to process transactional data, customer profiles, payment history, lending datasets, and analytical workloads.
- Designed enterprise **MLOps** workflows using **Amazon SageMaker**, **MLflow**, **Docker**, **Kubernetes**, **Jenkins**, and **GitHub Actions** to automate model deployment, versioning, monitoring, CI/CD pipelines, and production lifecycle management.
- Evaluated early **Generative AI** capabilities using **OpenAI APIs** and **Amazon Bedrock** for intelligent document summarization, financial knowledge retrieval, contextual search, and internal productivity workflow automation.
- Developed interactive **Power BI** and **Tableau** dashboards using **SQL** to visualize lending performance, portfolio health, credit risk metrics, operational KPIs, predictive analytics, executive business insights for portfolio performance analysis.
- Optimized production **Machine Learning models** through **feature engineering**, **hyperparameter tuning**, **cross-validation**, and **statistical evaluation**, improving **prediction accuracy**, **operational efficiency**, and **model reliability**.
- Developed secure cloud-native AI applications using **Amazon SageMaker**, **AWS Lambda**, **Amazon ECS**, **AWS IAM**, and **AWS Secrets Manager** to support scalable model deployment, secure API integration, and enterprise production workloads.

- Collaborated with data engineers, software developers, risk analysts, compliance teams, product owners, and business stakeholders using **Agile Scrum** to deliver scalable, production-ready **Applied AI** and **Machine Learning** solutions.

Environment: Python, SQL, FastAPI, Flask, Apache Spark, PySpark, Databricks, Scikit-learn, XGBoost, Gradient Boosting, Logistic Regression, MLflow, Amazon SageMaker, AWS Glue, Amazon S3, Snowflake, Power BI, Tableau, spaCy, NLTK, Hugging Face Transformers, AWS Textract, AWS Comprehend, OpenAI API, Amazon Bedrock, Docker, Kubernetes, Amazon ECS, AWS Lambda, Amazon API Gateway, AWS IAM, AWS Secrets Manager, GitHub Actions, Jenkins, REST APIs, SHAP, Agile Scrum.

Client: State of New York, Albany, NY

Machine Learning Engineer

Jun 2019 – Jan 2021

- Architected scalable **Machine Learning** solutions on **Google Cloud Platform (GCP)** using **Python** and statistical modeling for revenue forecasting, fraud detection, budget optimization, and enterprise financial planning.
- Engineered end-to-end **Machine Learning** pipelines using **Scikit-learn**, **TensorFlow**, **MLflow**, **Google AI Platform**, and **Cloud Storage** for feature engineering, automated model training, deployment, monitoring, and lifecycle management.
- Developed predictive **ML** models using **Random Forest**, **XGBoost**, and **Decision Trees** for classification, regression, anomaly detection, and financial forecasting, supporting budget planning, risk assessment, and strategic decision-making.
- Built scalable feature engineering pipelines for **Machine Learning** models using **Apache Spark**, **PySpark**, **Cloud Dataproc**, and **BigQuery** to transform large-scale financial datasets into optimized training data for predictive modeling.
- Implemented streaming data pipelines using **Apache Kafka**, **Spark Streaming**, and **Cloud Pub/Sub** to process financial transactions, operational events, and real-time analytics supporting forecasting and fraud detection.
- Developed NLP-based feature extraction pipelines using **spaCy**, **NLTK**, and **BERT** for document classification, named entity recognition, semantic analysis, and structured transformation of government reports to support predictive analytics.
- Operationalized production **ML** models using **Docker**, **Google Kubernetes Engine (GKE)**, **MLflow**, and **Cloud Build** for automated deployment, model versioning, scalable inference, continuous monitoring, and **CI/CD-driven MLOps** workflows.
- Engineered **ML** dataset preparation workflows using **BigQuery**, **Cloud SQL**, and **Snowflake** for large-scale data integration, transformation, dataset preparation, and preprocessing, improving query performance, training efficiency, and model readiness.
- Developed interactive **Power BI** and **Tableau** dashboards visualizing financial KPIs, revenue trends, expenditure forecasts, fraud indicators, **ML** model performance, predictive analytics, and executive reporting insights for business **stakeholders**.
- Optimized production **ML** models through feature engineering, hyperparameter tuning, cross-validation, statistical evaluation, and performance monitoring, improving forecasting accuracy, model reliability, and operational efficiency.
- Applied secure **Machine Learning** engineering practices using **IAM**, encryption, audit logging, secure data governance, access controls, and compliance monitoring to ensure regulatory compliance, enterprise security, and data protection.
- Collaborated with financial analysts, data engineers, project managers, and business stakeholders using **Agile Scrum** to deliver scalable, production-ready **Machine Learning** solutions supporting statewide financial operations.

Environment: Python, SQL, Scikit-learn, TensorFlow, Decision Trees, XGBoost, Random Forest, MLflow, Apache Spark, PySpark, Apache Kafka, Spark Streaming, spaCy, NLTK, BERT, Google Cloud Platform (GCP), Google AI Platform, BigQuery, Cloud Storage, Cloud Dataproc, Cloud Pub/Sub, Cloud SQL, Google Kubernetes Engine (GKE), Cloud Build, Docker, Git, Snowflake, Power BI, Tableau, Agile Scrum.

Client: TJ Maxx, Framingham, MA

Data Scientist

Oct 2016 – May 2019

- Developed predictive **ML** models using Python, Scikit-learn, Random Forest, Gradient Boosting, and Logistic Regression to improve demand forecasting, inventory optimization, pricing strategies, merchandising analytics, and retail sales performance.
- Performed exploratory data analysis, feature engineering, statistical modeling, hypothesis testing, and data visualization using **Pandas**, **NumPy**, and **SciPy** to uncover customer purchasing patterns, sales trends, and actionable business insights.
- Designed customer segmentation models using **K-Means Clustering**, **Hierarchical Clustering**, **PCA**, and behavioral analytics to improve targeted marketing campaigns, personalized promotions, customer retention, and merchandising effectiveness.
- Built recommendation systems using **collaborative filtering**, **association rule mining**, and **market basket analysis** to generate personalized product recommendations, cross-selling opportunities, customer affinity insights, and promotional optimization strategies.

- Developed time-series forecasting models using **ARIMA**, **Holt-Winters**, and statistical forecasting techniques to predict product demand, seasonal purchasing behavior, inventory replenishment requirements, and store-level sales performance.
- Evaluated predictive models using feature selection, **hyperparameter tuning**, **cross-validation**, **statistical validation**, and performance metrics to improve model **accuracy**, **robustness**, **reliability**, and **business decision-making**.
- Performed **A/B testing**, statistical experimentation, and regression analysis to evaluate pricing strategies, promotional campaigns, customer engagement, and merchandising effectiveness using quantitative business metrics.
- Developed interactive **Power BI** and **Tableau** dashboards visualizing merchandising KPIs, sales performance, customer purchasing behavior, inventory trends, demand forecasts, and executive insights for supporting strategic retail decisions.
- Collaborated with merchandising teams, pricing analysts, marketing managers, supply chain stakeholders, and executive leadership using **Agile Scrum** to deliver data-driven insights supporting enterprise retail strategy and operational excellence.

Environment: Python, SQL, Pandas, NumPy, SciPy, Scikit-learn, Random Forest, Gradient Boosting, Decision Trees, Logistic Regression, K-Means Clustering, Hierarchical Clustering, Principal Component Analysis (PCA), Association Rule Mining, Market Basket Analysis, ARIMA, Holt-Winters, Matplotlib, Power BI, Tableau, PostgreSQL, SQL Server, Git, Agile Scrum.

Client: Biocon, Hyderabad, India.

Python Full Stack Developer

Feb 2015- Aug 2016

- Designed enterprise web applications using Python, Django, Flask, HTML5, CSS3, JavaScript, jQuery, and Bootstrap to automate pharmaceutical workflows, inventory management, operational reporting, business process automation and workflow efficiency.
- Developed scalable backend modules using **Python**, **Django MVC**, and **REST APIs**, implementing reusable business logic, authentication, authorization, session management, and secure role-based access control for enterprise healthcare applications.
- Built responsive user interfaces using **HTML5**, **CSS3**, **Bootstrap**, **React.js**, **JavaScript**, **AJAX**, and **jQuery**, delivering interactive dashboards, reporting portals, data entry modules, and user-friendly enterprise web applications across business functions.
- Designed normalized relational database schemas using **MySQL**, **PostgreSQL**, **SQL Server**, and **SQL**, optimizing stored procedures, indexes, complex queries, database performance, and transaction processing for enterprise business applications.
- Developed automated data processing and reporting utilities using **Python**, **Pandas**, **SQL**, **Matplotlib**, and **Microsoft Excel**, improving operational reporting, reporting efficiency, data validation, and reporting accuracy across departments.
- Implemented server-side validation, exception handling, logging, application security, and performance optimization while supporting production deployments, defect resolution, system enhancements, and enterprise software maintenance activities.
- Collaborated with business analysts, QA engineers, project managers, and cross-functional development teams using **Git**, **Jenkins**, and **Agile Scrum** to deliver scalable, maintainable, and production-ready enterprise applications.

Environment: Python 2.7, Django, Flask, HTML5, CSS3, JavaScript(ES5), jQuery, Bootstrap, AJAX, REST APIs, SQL, MySQL, PostgreSQL, SQL Server, Pandas, Matplotlib, Microsoft Excel, Git, Jenkins, Linux, Apache HTTP Server, Agile Scrum.

Education: Bachelors in computer science, Guru Nanak Institute of Technology, India