****

Smart Phone with solid fill\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

**Yuvaraj Sriramoju**

**Generative Ai Engineer**

**EDUCATION**

**University of North Texas**, Master’s in

Artificial Intelligence, USA, from 2023 - 2025



Email with solid fill\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

**OBJECTIVE**

Innovative and results-driven AI Generative Engineer with 5+ years of experience in developing and deploying advanced generative models using Deep Learning, NLP, and Transformer architectures and leverage expertise in LLMs, GANs, and multimodal AI systems to design scalable.

**TECHNICAL SKILLS**

**Programming Languages**: Python, JavaScript, SQL

**Frameworks & Libraries**: PyTorch, TensorFlow, Hugging Face

Transformers, Kera’s, Lang

**Model Architectures**: GPT, BERT, T5, CLIP, Diffusion Models,

**Generative AI Tools**: OpenAI API, Anthropic Claude, Cohere, Hugging Face Hub, Stability AI, Dream Booth

**Prompt Engineering**: Few-shot, Zero-shot, Chain-of-Thought.

**NLP & LLMs**: Text generation, Summarization, Q&A,

**Vector Databases**: FAISS, Pinecone, Weaviate, Chroma DB

**Data Engineering**: Apache Airflow, Kafka, Spark, Big Query, Snowflake

**Model Deployment**: Docker, Kubernetes, Fast API, Flask, TensorRT, ONNX

**MLOps Tools**: MLflow, Weights & Biases, DVC, KServe, SageMaker, Vertex AI

**Visualization & UI**: Gradio, Streamlet, Dash, Jupyter Notebooks

**Cloud Platforms**: AWS, Azure, GCP

**DevOps & CI/CD**: Git, GitHub Actions, Terraform, Helm, Jenkins

**Monitoring & Logging**: Prometheus, Grafana, ELK Stack, Open Telemetry

**Security & Ethics**: Differential Privacy, Federated Learning, Responsible AI

**ML Ops Tools**: MLflow, DVC, Tensor

**Development Tools**: Jupyter, VSCode,

PyCharm, Collab

**Data Storage**: SQL, NoSQL, PostgreSQL, MongoDB, Redis, Elasticsearch

**Speech and Audio AI**: Whisper, Tacotron 2, Wave Glow, Google Speech API

**Reinforcement Learning**: OpenAI Gym, TensorFlow Agents

**PROFILE SUMMARY**

* **4+** years of experience in developing and deploying AI and ML solutions with specialization in **Generative AI**, including **LLMs**, **GANs**, **VAEs**, and **diffusion models** for text, image, and audio generation.
* Expertise in **Python**, **TensorFlow**, **PyTorch**, and **Kera’s** for building and training complex neural network architectures.
* Built and fine-tuned large language models using **Hugging Face Transformers**, **BERT**, **GPT**, and **T5**, enabling applications in chatbots, summarization, and code generation.
* Developed scalable ML pipelines using **Kubeflow**, **MLflow**, and **Airflow** to automate data preprocessing, training, evaluation, and deployment.
* Leveraged **AWS SageMaker**, **Azure ML Studio**, and **Google Vertex AI** for building, training, and serving ML models in cloud production environments and Integrated **OpenAI**, **Anthropic Claude**, and **Cohere** APIs to power intelligent enterprise solutions using foundation models for text and document understanding and Used **Docker**, **Kubernetes**, and **Terraform** for containerization, orchestration, and infrastructure automation in AI/ML workflows and hands-on experience with vector databases like **Pinecone**, **Weaviate**, and **FAISS** for implementing semantic search and retrieval-augmented generation (RAG) and Utilized **Neo4j**, **GraphQL**, and **RDF** for building knowledge graphs and integrating them with LLMs for context-aware reasoning and retrieval.
* Designed data architectures with **Apache Kafka**, **Spark**, and **Databricks** for real-time processing and large-scale model training.
* Collaborated in cross-functional Agile teams using **Git**, **JIRA**, and **Confluence**, maintaining CI/CD pipelines and ensuring model reproducibility and integrity and deployed serverless AI applications using **AWS Lambda**, **API Gateway**, and **DynamoDB**, reducing latency and operational costs in model inference and managed continuous integration workflows using **GitHub Actions**, **CircleCI**, and **Argo Workflows**, ensuring rapid experimentation and deployment of generative pipelines and developed custom generative pipelines using **Google Vertex AI**, **SageMaker JumpStart**, and **Azure OpenAI**, streamlining deployment across cloud platforms.
* Developed real-time chat interfaces using **Next.js**, **WebSocket’s**, and **Socket.IO** integrated with LLM backends for responsive AI assistants.
* Created voice and speech-to-text applications with **Whisper**, **Mozilla Deep Speech**, and **Assembly AI**, enhancing accessibility and AI-driven transcription.

**WORK EXPERIENCE**

**Generative Ai Engineer**

**Texas Instruments |** **Dallas, Texas, USA Sep 2024– Present**

**Objective: Texas Instruments Incorporated** is a global leader in **analog and embedded processing semiconductors**. TI designs, manufactures, tests. Designing and developing ML systems and identifying suitable data representation methods, analysing data quality, performing statistical analysis, and building suitable machine learning models and applications and finding trends in large data sets.

**Responsibilities:**

* Developed edge AI solutions using **TensorFlow Lite**, **ONNX Runtime**, and **NVIDIA Jetson** for deploying real-time vision and audio analytics models on TI’s embedded platforms.
* Built anomaly detection systems for semiconductor fabrication using **PyTorch**, **Scikit-learn**, and **XGBoost**, improving defect classification and predictive maintenance in wafer production.
* Engineered NLP-based knowledge assistants using **LangChain**, **OpenAI API**, and **FAISS** to provide real-time support for design engineers working with **analog** and **mixed-signal** IC documentation.
* Integrated **Apache Kafka**, **Apache Airflow**, and **Spark Structured Streaming** to enable real-time analytics pipelines across automated test equipment (ATE) and manufacturing data lakes.
* Utilized **Azure ML Studio**, **Docker**, and **Kubernetes** to deploy CI/CD pipelines for machine learning models applied to chip design optimization and yield improvement and applied **BERT**, **CLIP**, and **YOLOv5** models on silicon validation images and engineering reports to automate issue triage and accelerate root-cause analysis in post-silicon debug.
* Constructed digital twins of production lines using **Neo4j**, **Databricks**, and **Delta Lake**, enhancing traceability and root-cause diagnostics of component-level faults and designed and deployed **Gradio** and **Streamlit** applications for internal R&D use, enabling real-time interaction with simulation models and circuit-level performance predictors.
* Developed generative design tools with **Three.js**, **WebGPU**, and **AutoML**, enabling automated prototyping of PCB layouts and analog circuit topologies and leveraged **Azure Cognitive Services** and **Vertex AI** for building multilingual voice and chatbot interfaces that assist embedded system developers in TI's global operations.
* Integrated **Speech Brain**, **Coqui TTS**, and **Whisper** into accessibility interfaces for debugging embedded applications using voice commands on TI's evaluation modules and applied **Media Pipe**, **OpenCV**, and **gesture recognition** pipelines to build gesture-based HMI systems for automotive and industrial semiconductor demonstrations.
* Automated hyperparameter tuning and model tracking with **MLflow**, **Weights & Biases**, and **Google AutoML**, improving training workflows for silicon performance prediction models.
* Used **Python**, **Prometheus**, and **Grafana** to build telemetry dashboards for monitoring test equipment and yield analytics across global fab and assembly sites and employed **LabVIEW** and **NI Test Stand** to automate semiconductor device validation, improving test throughput for analog/mixed-signal products

**Environment:**

**Machine Learning Engineer**

**Charles Schwab |Westlake, Texas, USA Oct 2023 – Aug 2024**

**Objective: Charles Schwab Corporation** is a leading provider of **financial services,** specializing in **wealth management,** **securities brokerage, banking, asset management retirement planning**. As a machine learning engineer selecting appropriate data sets before performing data collection and data modelling and involves algorithms and statistical models that identify patterns, relationships, and insights within data and recognize patterns, and judge the new incoming data.

**Responsibilities:**

* Developed and deployed predictive models using **Python**, **TensorFlow**, and **Scikit-learn** to forecast market trends, client portfolio performance, and identify investment opportunities across Schwab’s brokerage and wealth management platforms and streamlined data ingestion from **trading platforms**, **CRM systems**, and **client financial records** using **Apache Kafka**, **Apache NiFi**, and **Airflow**, enabling real-time portfolio analytics and risk assessment.
* Processed large-scale financial datasets using **Hadoop**, **Apache Spark**, and **Delta Lake** to improve trading insights, client segmentation, and credit risk evaluation.
* Designed and maintained interactive dashboards with **Power BI**, **Tableau**, and **Looker**, providing advisors and executives with real-time views into asset allocation, AUM growth, and client engagement metrics.
* Deployed machine learning models for fraud detection and robo-advisory services using **AWS SageMaker**, **GCP Vertex AI**, and **Azure ML**, enhancing personalization and compliance.
* Automated CI/CD pipelines using **Jenkins**, **Terraform**, and **GitHub Actions** for robust model deployment and infrastructure-as-code implementation across Schwab’s data science ecosystem.
* Managed infrastructure for analytics workloads using **Docker**, **Kubernetes**, and **Ansible**, supporting scalable deployment of customer profiling and transaction monitoring models.
* Designed optimized data warehouses using **Snowflake**, **PostgreSQL**, and **MySQL** to store and serve client portfolio data, trade execution logs, and retirement account transactions, built real-time alerting systems integrated with **Apache Kafka**, **Fast API**, and **Redis** for risk threshold monitoring, insider trading detection, and market anomaly notifications.
* Leveraged **Bitbucket**, **Git**, and **GitLab** for version control, team collaboration, and secure audit trails across investment data science projects and engineered recommendation engines using **LightGBM**, **XGBoost**, and **Streamlit** to assist financial advisors in proposing suitable investment strategies based on client profiles.
* Integrated **dbt** and **Snow pipe** to automate data transformations and ingestion pipelines across Schwab’s investment, retirement, and banking services.

**Environment: Python, TensorFlow, Scikit-learn, PyTorch, Apache Kafka, Apache NiFi, Hadoop, Apache Spark, Tableau, Power BI, AWS SageMaker, Jenkins, CI/CD, Terraform, Ansible, Docker, Kubernetes, PostgreSQL, MongoDB, MySQL, Git, GitHub, Bitbucket,** Snowflake,Redis, Fast API, LightGBM, **Snow pipe.**

**Data Scientist**

**Deutsche Bank | Bangalore, India June 2022 – May 2023**

**Objective:** **Deutsche Bank** is a leading global **investment bank and financial services company**. Responsible to use statistical techniques, machine learning, and visualization tools to analyse large amounts of data to identify patterns and anomalies and identity theft, phishing, and payment card fraud and conducting quantitative analysis to obtain insights, developing these into workable solutions, and then following these through to successful delivery.

**Responsibilities:**

* Implemented **predictive analytics** to enhance **risk management**, **fraud detection**, and **credit scoring** systems for insurance claims processing, improving decision-making accuracy and operational efficiency.
* Designed and deployed automated reporting systems using **Tableau** and **Power BI** to visualize key performance indicators (KPIs) and provide actionable insights to senior management and stakeholders.
* Built and optimized **deep learning models** using **Keras**, **TensorFlow**, and **PyTorch** for complex data analysis tasks, enabling advanced claims prediction and customer behaviour analysis.
* Utilized **big data technologies** such as **Hadoop** and **Spark** to process and analyse large volumes of insurance claims and customer data efficiently, driving faster decision-making and claims processing.
* Leveraged **SQL** and **NoSQL databases** like **MySQL**, **MongoDB**, and **PostgreSQL** to extract, store, and manage massive datasets related to insurance claims, customer profiles, and policy details.
* Developed and implemented fraud detection systems using **anomaly detection** algorithms to identify and prevent fraudulent activities within insurance claims and enabled **real-time analytics** for live insurance claim monitoring and fraud detection using **Apache Kafka** and **Apache Flink**, ensuring timely response to emerging issues.
* Participated in designing robust **data pipelines** using **Apache Airflow** and optimized **ETL processes** to streamline data collection, transformation, and storage for improved data accessibility and analytics.
* Developed automated workflows using **Jenkins**, **GitLab CI**, and other **DevOps tools** to ensure continuous integration and delivery (CI/CD) of machine learning models, improving operational efficiency and deployment speed.
* Utilized **version control tools** such as **Git** and **GitHub** for collaborative code development, ensuring seamless code management and efficient collaboration across teams.
* Designed and implemented **multi-tier architectures** for high-volume transaction systems, using **Apache Cassandra** for distributed storage and retrieval of time-series insurance data, ensuring low-latency and high-availability solutions.
* Implemented **data lakes** and optimized **data warehousing** solutions using **Snowflake** and **Google Big Query** to store and analyse vast amounts of structured and unstructured insurance data, enhancing business intelligence capabilities.
* Applied **Natural Language Processing (NLP)** techniques to analyse customer feedback, claims data, and call centre transcripts using **spaCy** and **NLTK**, enabling sentiment analysis and customer satisfaction insights.
* Utilized **Docker** for containerizing machine learning models and **Kubernetes** for orchestrating and managing containerized applications, ensuring seamless scaling and deployment in production environments.
* Developed **real-time streaming analytics** for processing and analysing live transactional data with **Apache Storm** and **Apache Flink**, enabling quick decision-making for high-stakes insurance claims and Utilized **Elasticsearch** for fast search and analysis of large sets of insurance-related documents, claims, and policy data, improving data retrieval times and insights.

**Environment:** Tableau, Power BI, Kera’s, TensorFlow, PyTorch, Hadoop, Spark, SQL, NoSQL, MySQL, MongoDB, PostgreSQL, Apache Kafka, Apache Flink, Apache Airflow, ETL, Jenkins, GitLab CI, Git, GitHub,Apache Cassandra, Snowflake, Google Big Query, spaCy, NLTK, Apache Storm, Apache Flink, Elasticsearch.

**Data Scientist**

**Go Digit Insurance | Bangalore, India June 2021 – May 2022**

**Objective:** **Go Digit General Insurance Ltd.** is a **digital-first, full-stack general insurer** offers a diversified portfolio of 74 insurance product. Build and evaluate machine learning models for tasks like risk assessment, fraud detection, customer segmentation, and predictive pricing. They utilize techniques like regression, classification, clustering, and deep learning.

**Responsibilities:**

* Built real-time insurance claim processing and fraud detection pipelines using **Apache Kafka**, **Apache Flink**, and **Apache Pulsar**, enabling high-speed data ingestion and anomaly detection.
* Applied **ARIMA**, **Prophet**, and **SARIMA** models for time-series forecasting of claim frequency, hospitalization trends, and premium payment behaviours and created dynamic visualizations using **Tableau**, **Power BI**, **Matplotlib**, and **Seaborn** to present policy adoption, regional trends, and SLA performance.
* Processed and analysed large-scale claim datasets using **Apache Spark**, **Hadoop**, and **Delta Lake** to support actuarial modelling and reinsurance decisions and managed heterogeneous datasets with **MongoDB**, **Cassandra**, and **Elasticsearch**, enabling scalable handling of diagnostic records, EHRs, and claims metadata.
* Orchestrated automated ETL workflows using **Apache Airflow**, **Talend**, and **Apache NiFi**, integrating data from TPAs, hospital APIs, and internal insurance systems and deployed machine learning pipelines using **MLflow**, **Docker**, **Kubernetes**, and **Flask** to support real-time fraud scoring and hospitalization prediction.
* Integrated structured and unstructured data using **Snowflake**, **PostgreSQL**, **SQL**, and **NoSQL** to power analytics across underwriting, agent performance, and product recommendation and streamlined CI/CD operations with **Jenkins**, **Terraform**, and **GitHub Actions**, enabling continuous deployment of customer onboarding and claims triage services.
* Collaborated across teams using **Jupyter Notebooks**, **Google Colab**, **Confluence**, and **Jira**, ensuring transparency and reproducibility of ML workflows and built personalized insurance recommendation systems using **PyTorch**, **Kera’s**, **TensorFlow**, and **XGBoost** based on lifestyle data and medical history.
* Engineered wellness risk stratification models using **LightGBM**, **Cat Boost**, and **scikit-learn**, improving policy targeting and premium optimization and developed NLP solutions using **spaCy**, **BERT**, **Hugging Face Transformers**, and **LangChain** for document summarization, chatbot interactions, and RAG-based policy queries.
* Designed real-time visualizations and efficiency metrics using **D3.js**, **Plotly**, and **Dash**, supporting executive decision-making in hospital partnerships and integrated cloud-native services via **GCP**, **AWS**, for deploying scalable AI-driven claim advisory tools and customer experience solutions.
* Implemented **GraphQL** APIs to streamline customer onboarding and policy lookup processes, reducing response latency and simplifying front-end integrations and employed **Neo4j** for graph-based relationship mapping between customers, agents, and claims to detect fraudulent clusters and referral loops and used **Prometheus**, **Grafana**, and **ELK Stack** to monitor real-time performance metrics of data pipelines and application services.
* Secured API gateways using **Kong**, **Ambassador**, and **APIgee**, ensuring scalability and observability across digital insurance channels.

**Environment: Apache Kafka, Apache Flink, Apache Pulsar, ARIMA, Prophet, SARIMA, Apache Spark, Hadoop, MongoDB, Cassandra, Apache Airflow, Flask, MLflow, Docker, Kubernetes, Snowflake, SQL, PostgreSQL, NoSQL, Jenkins, Terraform, PyTorch, Kera’s,** XGBoost, LightGBM, Cat Boost, Spacey, BERT, D3.js, Plotly, Seaborn, Talend, Apache NiFi, **GCP, AWS,** Neo4j, Prometheus, Grafana, ELK Stack, Kong, Ambassador, APIgee