Name: ACHYUTHA P
Role: Generative AI & Machine Learning Engineer
Email address: achyuthapranavi268@gmail.com
LinkedIn: http://www.linkedin.com/in/achyutha-p-89a3b5216
Contact: +1 5513614995

PROFESSIONAL SUMMARY

- Senior Generative AI & Machine Learning Engineer with **12+ years** of hands-on experience building and deploying **AI and data-driven systems** across finance, healthcare, retail, insurance and telecom domains. Specialized in **LLM-powered applications, RAG, model fine-tuning**, prompt engineering and scalable cloud-based ML deployments, with **2 years of focused enterprise GenAI experience** in production environments since 2023.
- Built and deployed enterprise-grade Generative AI applications using **Amazon Bedrock** and **Azure OpenAI**, implementing **RAG pipelines** and multi-agent workflows using **LangChain and LangGraph** for document intelligence and knowledge retrieval across regulated financial and clinical environments.
- Engineered multi-agent LLM workflows and RAG pipelines integrating **Amazon OpenSearch, Azure Cognitive Search, FAISS** and **Pinecone**, applying hybrid BM25 and vector retrieval for domain-specific financial research and healthcare analytics use cases.
- Developed ML pipelines using **AWS SageMaker, Vertex AI, Azure ML, Databricks** and **MLflow** to support model training, evaluation, deployment and monitoring across multi-cloud environments.
- Developed and fine-tuned transformer-based NLP models including **BERT, GPT, LLaMA, Mistral** using **PEFT techniques (LoRA/QLoRA)** for financial document analysis, clinical text processing, sentiment analysis and entity extraction in high-stakes regulated environments.
- Applied time-series forecasting, ensemble learning and deep learning using **TensorFlow, PyTorch, XGBoost, ARIMA, Prophet** and **LSTM** to support demand prediction, risk assessment, fraud detection and pricing optimization across large-scale production environments.
- Developed scalable data pipelines using **PySpark, AWS Glue, Google Cloud Dataflow, Azure Data Factory, Delta Lake, Snowflake** and **dbt** to process structured and unstructured data for ML model training and real-time inference.
- Integrated explainability techniques using **SHAP** and **LIME** to enhance model transparency and meet compliance and auditability requirements in HIPAA and financial regulatory environments.
- Deployed production ML and GenAI services using **Docker, Amazon EKS, GKE** and **Azure Kubernetes Service**, supporting scalable real-time inference and monitoring using Amazon CloudWatch, Google Cloud Monitoring, Azure Monitor and Prometheus.
- Applied knowledge graph techniques using **Neo4j** and **NetworkX** to enable relationship-aware fraud detection and compliance analytics with explainable, traceable outputs.
- Built modular LLM application layers integrating retrieval, prompt orchestration, model inference and response validation to deliver auditable, enterprise-ready GenAI services.
- Collaborated with data engineering, DevOps and product teams to deliver production-ready AI and Generative AI solutions aligned with regulatory, compliance and business requirements.

TECHNICAL SKILLS

Programming	Python, SQL, PySpark, SAS, Java, Bash
Generative AI & LLM	Amazon Bedrock (Claude/Titan), Azure OpenAI (GPT-4), LangChain, LangGraph, Semantic Kernel, Hugging Face Transformers, PEFT (LoRA/QLoRA), MedLM, RAG Architectures
Vector Databases & Search	FAISS, ChromaDB, Pinecone, Amazon OpenSearch, Azure Cognitive Search, Hybrid Retrieval (BM25 + Vector), Semantic Search
ML Frameworks	PyTorch, TensorFlow, scikit-learn, XGBoost, Random Forest, Gradient Boosting
Cloud Platforms & AI/ML Services	AWS (core production experience): SageMaker, EMR, Redshift, EKS, OpenSearch, Comprehend, Textract, CloudWatch, Lambda, EC2 GCP (HCA projects): Vertex AI, BigQuery, Cloud Dataflow, Dataproc, GKE, Cloud Build, Document AI, Cloud Healthcare API, Cloud Monitoring, Looker Studio Azure: Azure ML, Azure Databricks, Azure Synapse Analytics, Azure Data Factory, Azure Functions, AKS, Azure Monitor, Azure DevOps Data warehousing / processing: Snowflake, Delta Lake
MLOps & DevOps	MLflow, Docker, Kubernetes (AKS, EKS, GKE), Azure DevOps, Git, Jenkins, CI/CD Pipelines, Model Monitoring, Drift Detection

NLP & Transformers	BERT, GPT, LLaMA, Mistral, spaCy, Text Summarization, Sentiment Analysis, Entity Recognition, Clinical NLP
Time-Series & Forecasting	ARIMA, Prophet, LSTM
Knowledge Graphs	Neo4j, NetworkX, Graph Embeddings, Relationship Modeling
Big Data & ETL	PySpark, pandas, AWS Glue, Google Cloud Dataflow, Delta Lake, dbt, Alteryx, Apache Airflow, Azure Data Factory
Data Warehousing	Amazon Redshift, Snowflake, BigQuery, SQL Server, MySQL, Oracle, Query Optimization
Data Visualization & BI	Power BI, Tableau, Amazon QuickSight, Matplotlib, Seaborn, Looker Studio
Responsible AI	SHAP, LIME, Amazon Bedrock Guardrails, Azure AI Content Safety
Statistical Analysis	A/B Testing, Hypothesis Testing, Cohort Analysis, Customer Segmentation, GLMs
Monitoring & Observability	Prometheus, AWS CloudWatch, Google Cloud Monitoring, Azure Monitor, Application Insights
APIs & Deployment	FastAPI, OAuth2, RBAC, REST APIs, Real-Time Inference, Batch Scoring, Event-Driven Architectures

WORK EXPERIENCE

GenAI/ML Engineer - Jefferies Financial Group Inc, New York, NY
May 2024 - Present

Building enterprise LLM applications, multi-agent workflows and scalable ML systems for financial research and risk analytics on AWS.

- Built and deployed AI/ML and Generative AI applications on AWS using **Amazon SageMaker, Amazon EMR** and **Amazon Redshift** to support secure and scalable model training and inference for financial research and risk analytics use cases.
- Developed reproducible ML pipelines using **Python, Amazon SageMaker** and **MLflow** to support experimentation tracking, CI/CD integration and consistent production deployment across model lifecycle stages.
- Deployed **LLM-powered enterprise assistants** using **Amazon Bedrock** (Claude/Titan) and **LangChain** to automate financial document analysis, research summarization and knowledge retrieval, significantly reducing manual search and review effort for analysts.
- Engineered **multi-agent Generative AI workflows** using **LangGraph** and **LangChain Agents** to automate ingestion, retrieval and validation of SEC filings and financial research reports, improving workflow efficiency by approximately 33%.
- Developed **Retrieval-Augmented Generation pipelines** using **Amazon OpenSearch** with hybrid BM25 and vector retrieval, along with **FAISS** and **Pinecone**, implementing hierarchical chunking, semantic re-ranking and contextual filtering to reduce hallucinations by approximately 28% and improve citation accuracy.
- Fine-tuned domain-adapted LLMs using **PEFT (LoRA/QLoRA)** with **PyTorch** and **Hugging Face Transformers**, improving financial summarization and risk entity extraction accuracy by approximately 14% on internal evaluation benchmarks.
- Implemented query classification and routing logic to direct complex financial queries to **Amazon Bedrock (Claude 3)** and routine requests to cost-optimized open-source models, reducing average inference cost by approximately 20%.
- Built **LLM-powered analyst copilots** supporting earnings analysis, compliance workflows and policy navigation using **LangChain**, Amazon Bedrock and vector search, accelerating analyst onboarding and improving daily productivity.
- Deployed ML and GenAI workloads on **Amazon SageMaker** and **Amazon EKS** with automated monitoring and retraining workflows, ensuring consistent production inference at scale.
- Developed production-grade **GenAI APIs** using **FastAPI** with OAuth2, RBAC and asynchronous token streaming, enabling secure integration into internal financial systems and analyst platforms.
- Implemented real-time inference pipelines using event-driven architectures, reducing prediction latency by approximately 35% for trading and risk analysis workflows.
- Integrated a **Neo4j** knowledge graph layer for relationship-aware fraud detection and compliance risk scoring, enabling explainable answers with traceable counterparty and transaction paths.
- Applied **Amazon Bedrock Guardrails, SHAP** and LIME to meet explainability, auditability and regulatory compliance requirements for AI-generated financial outputs.
- Built and maintained **LLM evaluation pipelines** combining automated metrics and human-in-the-loop validation to benchmark Amazon Bedrock foundation models against fine-tuned open-source alternatives.
- Built production monitoring and observability systems using **Amazon CloudWatch, AWS X-Ray** and **Prometheus**, reducing MTTR by approximately 30% and improving reliability of AI services.

Technologies: Amazon Bedrock (Claude 3/Titan), LangChain, LangGraph, Amazon OpenSearch, FAISS, Pinecone, Hugging Face Transformers, PyTorch, PEFT (LoRA/QLoRA), Python, FastAPI, Docker, Amazon EKS, Amazon SageMaker, Amazon EMR, Amazon Redshift, Amazon Comprehend, AWS Textract, PySpark, MLflow, CloudWatch, AWS X-Ray, Prometheus, SHAP, Amazon Bedrock Guardrails, Neo4j

---

AI/ML Engineer & Data Scientist - HCA Healthcare Inc., Nashville, TN
November 2022 - April 2024

Led ML, NLP and GenAI initiatives for clinical risk prediction, documentation automation and large-scale healthcare analytics on GCP. ML and NLP work from November 2022; GenAI from mid-2023 following enterprise availability of LLM platforms.

- Developed and deployed production **patient risk prediction models** using **Python, XGBoost, Random Forest** and **Google Cloud Vertex AI**, improving early-risk identification by approximately 22% and supporting proactive care delivery across multiple hospital sites.
- Built end-to-end **ML pipelines** using **Vertex AI, MLflow** and Cloud Build CI/CD, reducing model release cycle time by approximately 25% and enabling faster iteration across clinical modeling workflows.
- Built scalable **data ingestion and ETL pipelines** using **Cloud Dataflow, Google Cloud Storage** and **BigQuery** to centralize clinical data availability and accelerate downstream model development.
- Engineered **claims fraud and anomaly detection models** using ensemble learning techniques on clinical billing data, reducing false positives by approximately 30% and improving investigation prioritization efficiency for revenue cycle operations.
- Developed distributed data processing pipelines using **PySpark**, SQL and **Google Dataproc**, improving large-scale analytics performance by approximately 35% over legacy batch systems.
- Applied deep learning models using **TensorFlow** and **PyTorch** for clinical classification, anomaly detection and outcome forecasting, improving model robustness across multiple retraining cycles.
- Deployed low-latency inference endpoints on **Vertex AI** for real-time patient risk scoring, reducing prediction response time by approximately 40% for clinical decision support workflows.
- Implemented **explainable AI frameworks** using **SHAP** and LIME, reducing clinician review time by approximately 20% while maintaining HIPAA-compliant audit transparency.
- Integrated external social determinant of health datasets through advanced feature engineering and data fusion, enhancing patient outcome modeling and risk stratification accuracy.
- Built model monitoring and drift detection dashboards using Looker Studio and **Google Cloud Monitoring** to track prediction quality and automatically trigger retraining workflows when performance degraded.
- Developed **NLP pipelines** for clinical note analysis using **BERT, spaCy** and **Hugging Face Transformers**, improving structured data extraction accuracy from unstructured physician documentation.
- Implemented OCR and document intelligence workflows using **Google Cloud Document AI** and Vision API, reducing manual claims review effort across revenue cycle operations.
- Containerized and deployed ML services using **Docker** and **GKE** to support scalable, reliable inference across distributed clinical systems.
- From mid-2023, developed **GenAI-powered clinical documentation workflows** using **Vertex AI** and **MedLM (Med-PaLM 2)** on Google Cloud, supporting real-time ambient medical note generation from clinician-patient conversations across emergency department sites and integrating with EHR systems.

Technologies: Python, SQL, PySpark, XGBoost, Random Forest, TensorFlow, PyTorch, BERT, Hugging Face Transformers, spaCy, Vertex AI, BigQuery, Cloud Dataflow, Dataproc, Cloud Storage, Document AI, Vision API, Cloud Healthcare API, GKE, Cloud Build, Cloud Monitoring, MLflow, Docker, Looker Studio, SHAP, LIME, LangChain, Med-PaLM 2

AI/ML Engineer - Target Corp, Minneapolis, MN
January 2019 - October 2022

Developed forecasting, pricing and demand-planning ML systems across large-scale retail datasets spanning 1,800+ store locations.

- Developed scalable **ML-ready data pipelines** using **Python, PySpark** and Azure services to support demand forecasting and pricing optimization across retail merchandise categories.
- Built advanced **feature engineering frameworks** using **SQL, dbt** and **Snowflake** to generate time-series, customer-level and product-level features for downstream ML model training.
- Developed and deployed **demand forecasting models** using ARIMA, Prophet, **LSTM, TensorFlow** and **PyTorch**, improving forecast accuracy by approximately 25% on high-volume and seasonal SKUs.

- Implemented robust **model evaluation and backtesting frameworks** using RMSE, MAPE and rolling-window validation to ensure statistical and deep learning models met production performance thresholds.
- Applied ensemble learning using **XGBoost** and Random Forest to optimize pricing, promotion planning and markdown strategies across retail categories, resulting in improved margin outcomes.
- Integrated external market signals including Nielsen data, promotional calendars, competitive pricing and social sentiment into model features, improving responsiveness to seasonality and promotional demand shifts.
- Operationalized ML models using **Azure Functions** and CI/CD pipelines, enabling consistent batch and scheduled inference across 1,800+ retail locations nationwide.
- Developed optimized **ETL workflows** using **Azure Data Factory**, Alteryx and Python to automate data preparation and reduce manual processing effort by approximately 40%.
- Implemented data and model governance controls using **Azure Purview**, dataset versioning and access policies to ensure reproducibility and auditability of ML workflows.
- Built monitoring and alerting systems using **Azure Monitor** to track pipeline health and enable proactive issue detection before downstream business impact occurred.
- Conducted **A/B testing**, uplift modeling and cohort analysis, improving campaign targeting effectiveness by approximately 23%.
- Delivered **ML-driven forecasting dashboards** using **Power BI, Tableau** and Azure Synapse Analytics, surfacing forecast confidence intervals and risk indicators to merchandising and supply-chain teams.
- Collaborated with merchandising, supply-chain and marketing teams to translate ML forecasts into actionable pricing and inventory decisions, reducing overstock and improving planning accuracy.
- Established version-controlled ML workflows using **Git** and **Azure DevOps**, improving team reproducibility and standardizing model deployment practices.

Technologies: Python, PySpark, SQL, Azure Data Factory, Synapse Analytics, Data Lake Storage, Azure Functions, Azure Monitor, Azure Purview, Azure DevOps, Snowflake, dbt, TensorFlow, PyTorch, XGBoost, Random Forest, ARIMA, Prophet, LSTM, Tableau, Power BI, Alteryx, pandas, NumPy, Git, Jenkins

---

Machine Learning Engineer - Allstate Insurance, Northbrook, IL
September 2015 - December 2018

Built ML models for fraud detection, actuarial risk scoring and customer analytics in a regulated insurance environment.

- Developed and deployed supervised ML models using **Python, scikit-learn, Random Forest**, Gradient Boosting and Logistic Regression for customer churn prediction, fraud detection and insurance risk assessment across policyholder and claims datasets.
- Built **fraud detection and risk classification models** that improved detection accuracy by approximately 18%, reducing false positives and directly improving investigative efficiency for the claims operations team.
- Developed scalable **feature engineering pipelines** using **SQL** and **Python** to extract policyholder and claims-level features from large insurance databases.
- Developed **actuarial pricing and risk scoring models** using **GLMs** and statistical regression techniques to support underwriting decisions and optimize premium pricing across auto, home and life insurance lines.
- Built automated **ETL pipelines** using **AWS Glue** (adopted from 2017), Python and SQL to ingest and transform large-scale insurance datasets for model training and batch scoring.
- Trained and validated ensemble models (**XGBoost, Gradient Boosting, Random Forest**) using cross-validation, hyperparameter tuning and holdout testing to ensure generalization before production deployment.
- Implemented comprehensive **model evaluation frameworks** tracking precision, recall, AUC-ROC and calibration metrics before deployment into underwriting and claims scoring systems.
- Deployed ML models to production using **AWS Lambda** and EC2 with batch scoring workflows, enabling automated risk scoring across claims operations and underwriting processes.
- Built model monitoring and drift detection scripts using Python and SQL to track prediction distributions and feature stability across production environments.
- Applied statistical modeling using **SAS** (PROC LOGISTIC, PROC REG, PROC GENMOD) for actuarial analysis, renewal likelihood modeling and retention forecasting.
- Optimized large-scale feature extraction queries on **Amazon Redshift**, reducing data processing time by approximately 30% for model development workflows.
- Implemented **customer segmentation models** using K-means clustering and dimensionality reduction to support retention strategy and cross-sell campaign targeting.
- Collaborated with actuarial, underwriting and data engineering teams to translate regulatory and business requirements into deployable, compliant ML solutions within a regulated enterprise environment.

Technologies: Python, SQL, SAS, scikit-learn, pandas, NumPy, XGBoost, Random Forest, Gradient Boosting, Logistic Regression, K-means, AWS S3, Redshift, Athena, Glue, Lambda, EC2, CloudWatch

Python Developer / Data Analyst - Ooma Inc, Sunnyvale, CA
February 2013 - August 2015

Built Python-based data engineering and analytics systems supporting telecom operations and customer insights.

- Developed and maintained scalable **Python-based data processing systems** to ingest, transform and standardize telecom usage, call quality and customer interaction data across SQL Server, MySQL and Oracle databases.
- Built modular **ETL pipelines** using Python and SQL to automate recurring data workflows, reducing manual preparation effort across engineering and operations teams.
- Built internal analytics services and reporting pipelines supporting churn analysis, service reliability tracking and operational KPIs used by engineering and operations teams.
- Developed **predictive churn and call-drop models** using Logistic Regression and Decision Trees, enabling proactive retention and network optimization strategies.
- Implemented data validation and reconciliation scripts in **Python**, improving data accuracy by approximately 30% and ensuring cross-system consistency.
- Optimized **SQL** queries and indexing strategies to improve analytical query performance and dashboard responsiveness across large telecom datasets.
- Built and executed **A/B testing frameworks** to evaluate feature rollouts and service improvements, supporting data-driven product decisions.
- Conducted **VoIP and call-quality analytics** to identify network bottlenecks and latency issues, providing actionable infrastructure planning recommendations to engineering teams.
- Built executive and operational dashboards using **Tableau** and **Power BI**, translating technical metrics into business-impact insights for leadership teams.
- Collaborated with engineering and operations teams to integrate analytics outputs into production monitoring workflows and customer support processes.
- Maintained documentation for ETL logic, data models and reporting definitions to support knowledge transfer and long-term system maintainability.

Technologies: Python, R, SQL (SQL Server, MySQL, Oracle), pandas, NumPy, Logistic Regression, Decision Trees, Tableau, Power BI, ETL Pipelines

KEY PROJECTS

LLM Application & RAG Platform - Financial Domain (Jefferies, 2024-Present)

- Technologies: Amazon Bedrock (Claude 3 / Titan), LangChain, LangGraph, OpenSearch, FAISS, Pinecone, SageMaker, EKS, MLflow, FastAPI
- Hybrid retrieval architecture combining BM25 and vector search for grounded financial document responses
- Parameter-efficient fine-tuning using LoRA / QLoRA with PyTorch and Hugging Face Transformers
- Multi-agent orchestration pipelines for ingestion, retrieval validation and structured synthesis
- Embedding pipelines with domain-specific chunking and semantic re-ranking
- Secure FastAPI inference layer with OAuth2 and RBAC controls
- Containerized LLM workloads deployed on EKS with autoscaling policies
- LLM evaluation pipelines with automated benchmarking and human validation
- Observability dashboards tracking latency, token usage, drift and failure rates

Clinical GenAI & Risk Modeling Platform - Healthcare (HCA Healthcare, mid 2023-2024)

- Technologies: Vertex AI, MedLM (Med-PaLM 2), BigQuery, Cloud Dataflow, GKE, TensorFlow, PyTorch, MLflow
- Real-time clinical documentation assistant using MedLM-based ambient note summarization from clinician-patient conversations
- Patient risk prediction models using ensemble learning and deep neural networks for early clinical intervention
- Distributed ETL pipelines processing structured EHR data and unstructured clinical notes at scale
- Feature engineering workflows integrating social determinant of health datasets into risk stratification models
- Explainability pipelines using SHAP for model transparency and HIPAA-compliant clinical auditability
- Containerized inference services deployed on GKE with high-availability configuration
- Model lifecycle management with MLflow experiment tracking, versioning and artifact management
- Drift monitoring and performance dashboards using Google Cloud Monitoring to trigger automated retraining

EDUCATIONAL DETAILS

Master of Science in Computer Science
University of Central Missouri (August 2011 - January 2013)

Bachelor of Technology in Computer Science
Lovely Professional university (August 2007 - June 2011)

CERTIFICATIONS

- AWS Certified Machine Learning - Specialty
- AWS Certified Generative AI Developer - Professional
- Microsoft Certified: Azure AI Engineer Associate
- Google Cloud Professional ML Engineer
- PyTorch Developer Certification