ACHYUTHA PRANAVI K

achyutha2949@gmail.com || +1 201 430 5523 || LinkedIn

PROFESSIONAL SUMMARY

- Delivered production-grade **Generative AI systems** and **LLM-powered assistants** across semiconductor, automotive, manufacturing, and enterprise engineering domains, enabling documentation intelligence, knowledge retrieval, and decision-support workflows.
- Built and scaled enterprise **RAG platforms** using **Azure OpenAI (GPT-4)**, **LangChain**, and **LangGraph**, providing grounded, citation-backed responses over large technical corpora under strict latency and cost constraints.
- Implemented **multi-agent GenAI workflows** (planner, retriever, validator, compliance) with **LangGraph**, automating ingestion, summarization, reasoning, and validation pipelines and improving workflow throughput by ~35% with auditability.
- Fine-tuned and optimized open-source LLMs (**LLaMA, Mistral**) using **PEFT (LoRA / QLoRA)** for domain-specific summarization, classification, and reasoning, achieving ~15% accuracy gains with reduced compute overhead.
- Engineered **prompt engineering and instruction-tuning strategies**, including dynamic prompt routing and temperature control, to balance response quality, safety, and inference cost across enterprise GenAI workloads.
- Built production-grade **RAG systems** integrating **Azure Cognitive Search (BM25 + vector)** with **FAISS**, **Pinecone**, and **ChromaDB**, reducing hallucinations by ~30% while improving retrieval precision and traceability.
- Developed secure, scalable **GenAI inference services** using **FastAPI** and RESTful APIs with **OAuth2/RBAC**, structured logging, request tracing, and streaming responses for real-time and batch workloads.
- Containerized and deployed **GenAI and ML services** using **Docker** and CI/CD pipelines, with exposure to **Kubernetes-based deployments**, enabling reliable scaling and controlled production rollouts.
- Designed and implemented **ML and GenAI pipelines** spanning data ingestion, feature engineering, model development, evaluation, and deployment using **PySpark** and cloud-native data platforms.
- Established **LLM evaluation and benchmarking frameworks** combining automated metrics (**ROUGE, BLEU, Retrieval@k**) and human-in-the-loop review to assess accuracy, latency, grounding quality, and cost.
- Implemented **Responsible AI and safety controls** including content moderation, prompt guardrails, explainability (**SHAP, LIME**), and audit logging to support compliant enterprise adoption.
- Built **multi-turn conversational memory** and session-level state management for GenAI assistants, improving dialogue continuity, contextual recall, and reducing redundant interactions.
- Integrated **GenAI systems** into enterprise applications and internal platforms, enabling AI-powered copilots, auto-drafting workflows, and intelligent search aligned with business processes.
- Engineered scalable **data ingestion and enrichment pipelines** using **PySpark** and cloud data services to process structured telemetry, logs, and unstructured documents for downstream ML and GenAI use cases.
- Designed **observability and monitoring dashboards** exposing GenAI usage, latency, cost, reliability, and error patterns, reducing MTTR by ~30% and improving operational visibility.
- Collaborated closely with **engineering, data science, platform, and operations teams** to productionize GenAI solutions aligned with security, compliance, and real-world deployment constraints.
- Progressively transitioned from **predictive analytics and ML** into hands-on delivery of **Generative AI systems**, contributing across design, optimization, evaluation, deployment, and reliability.
- Demonstrated consistent delivery of **scalable, business-aligned AI systems** that balance accuracy, performance, cost efficiency, safety, and long-term maintainability.

TECHNICAL SKILLS

Languages	Python, SQL, JavaScript, Bash, R
Generative AI & LLMs	GPT-4, Azure OpenAI, OpenAI APIs, LLaMA, Mistral, Claude, Cohere, prompt engineering, few-shot prompting, instruction tuning, structured prompts, chain-of-thought prompting, LLM orchestration, prompt routing, safety-aware generation
Agentic AI & Orchestration	LangChain, LangGraph, LlamaIndex, multi-agent workflows, planner-retriever-validator patterns, tool calling, stateful workflows, multi-step reasoning
RAG & Vector Retrieval	Azure Cognitive Search (BM25 + vector), FAISS, Pinecone, ChromaDB, hybrid retrieval, chunking strategies, semantic similarity, re-ranking, citation grounding, context filtering, retrieval optimization
LLM Fine-Tuning & Optimization	PEFT, LoRA, QLoRA, task-specific fine-tuning, instruction-tuned models, evaluation-driven model selection
Machine Learning & Deep Learning	PyTorch, TensorFlow, Keras, scikit-learn, XGBoost, LightGBM, CatBoost, supervised learning, ensemble models
NLP & Text Analytics	BERT, Hugging Face Transformers, spaCy, named entity recognition, document classification, summarization, sentiment analysis, text normalization, preprocessing pipelines
Time-Series & Forecasting	ARIMA, Prophet, LSTM, demand forecasting, trend analysis, anomaly detection
MLOps, APIs & Deployment	FastAPI, RESTful APIs, Docker, CI/CD pipelines, MLflow, model versioning, batch inference, near-real-time inference, A/B testing, basic drift detection, Kubernetes (AKS)
Cloud & Data Engineering	Microsoft Azure, AWS, Azure Databricks, Azure Data Factory, AWS S3, AWS Lambda, EC2, Redshift, Glue, Athena, EMR, PySpark, Spark, Delta Lake, Airflow, Kafka
Databases & Data Stores	PostgreSQL, MySQL, Azure Synapse, Amazon Redshift, SQL Server, MongoDB, Cassandra

Document AI & Multimodal Systems	Azure Vision, OCR pipelines, OpenCV, document intelligence workflows, structured data extraction
Monitoring, Evaluation & Responsible AI	Azure Monitor, Application Insights, ROUGE, BLEU, retrieval@k, human-in-the-loop evaluation, SHAP, LIME, content moderation, prompt guardrails, responsible AI practices

PROFESSIONAL EXPERIENCE

Gen AI engineer **May 2024-Present**

Intel, New York

Enterprise Generative AI Platforms – Silicon, Manufacturing & Internal Engineering Systems

Responsibilities:

- Designed and implemented **LLM-powered engineering assistants** using **Azure OpenAI (GPT-4)**, **LangChain**, and **LangGraph** to support silicon design documentation search, manufacturing knowledge retrieval, and internal engineering workflows, reducing manual lookup and review effort.
- Implemented **multi-step, agent-based GenAI workflows** with **LangGraph** to automate ingestion, summarization, validation, and reasoning over technical specifications, manufacturing logs, and internal wikis, improving workflow throughput by ~35%.
- Built production-grade **RAG pipelines** using **Azure Cognitive Search (BM25 + vector)**, **FAISS**, **ChromaDB**, and **Pinecone** to query hardware specifications, firmware documentation, APIs, and SOPs, reducing hallucinations by ~30% and improving citation accuracy.
- Applied **query-time context filtering**, document chunking strategies, semantic similarity scoring, and lightweight re-ranking to improve retrieval precision across large technical corpora while controlling token usage and inference latency.
- Fine-tuned open-source LLMs (**LLaMA, Mistral**) using **PyTorch PEFT (LoRA / QLoRA)** for technical summarization, component classification, and log interpretation, achieving ~15% accuracy improvement on internal evaluation benchmarks.
- Implemented **prompt routing and model-selection logic** to route complex or safety-sensitive queries to **GPT-4**, while serving routine engineering queries with optimized open-source models, reducing overall inference cost by ~20%.
- Developed **LLM-powered engineering copilots** using **LangChain** and **Azure OpenAI** to assist with firmware development, driver APIs, system configuration, and internal knowledge discovery, improving onboarding efficiency and developer productivity.
- Built secure **GenAI inference APIs** using **FastAPI** with **OAuth2/RBAC**, structured logging, request tracing, and streaming responses to integrate LLM capabilities into internal engineering platforms.
- Containerized **GenAI services** using **Docker** with exposure to **Azure Kubernetes Service (AKS)**, enabling scalable batch and near–real-time inference with safe rollout patterns such as blue-green and canary releases.
- Integrated **vector stores and metadata layers** using **Azure Cognitive Search** and **Delta Lake** to support traceability, reproducibility, and controlled access to AI-assisted engineering workflows.
- Engineered **data ingestion and enrichment pipelines** using **PySpark**, **pandas**, and **Azure Data Factory** to process structured telemetry, system logs, and unstructured technical documents for downstream GenAI applications.
- Implemented **Responsible AI controls** using **Azure AI Content Safety**, prompt guardrails, system constraints, and explainability techniques to support safe, reliable, and policy-compliant AI behavior.
- Designed **LLM evaluation and benchmarking workflows** combining automated metrics and human-in-the-loop review to measure response accuracy, latency, cost, and grounding quality across **GPT-4** and open-source models.
- Developed **multi-turn conversational memory** with session-level state management to support long-running engineering troubleshooting and diagnostic conversations.
- Built **observability and feedback tooling** exposing retrieved context, citations, confidence indicators, and latency metrics, enabling rapid iteration and continuous improvement of GenAI systems.
- Delivered **Power BI dashboards** visualizing GenAI adoption, usage patterns, latency, cost, and reliability metrics for technical leads and engineering stakeholders.
- Integrated **monitoring and alerting** using **Azure Monitor** and **Application Insights**, reducing mean time to resolution (MTTR) by ~30% and maintaining high service reliability.
- Worked closely with **senior engineers and architects** on model selection, retrieval strategies, safety controls, and deployment patterns for enterprise GenAI systems.
- Collaborated with **silicon engineering, manufacturing, platform, data science, and IT teams** to productionize GenAI capabilities and support enterprise-wide adoption.

Environment:

GPT-4, LLaMA, Mistral, LangChain, LangGraph, Azure OpenAI, Azure Cognitive Search, FAISS, Pinecone, ChromaDB, FastAPI, Docker, Azure Kubernetes Service (AKS), PyTorch (LoRA / QLoRA), PySpark, Delta Lake, Azure Data Factory, Power BI, Azure Monitor, Application Insights

AI/ML Engineer / Data Scientist **Nov 2021 – Apr 2024**

General Motors., Detroit, MI

Fraud Detection, Warranty Analytics & Enterprise Risk Platforms

Responsibilities:

- Designed and deployed **supervised fraud detection and risk-scoring systems** using **Python**, **XGBoost**, **Random Forest**, and **Azure ML** to identify anomalous warranty claims and dealer fraud patterns, reducing false positives by ~18% and improving investigator throughput.
- Engineered **ensemble learning approaches** combining gradient boosting and tree-based models to stabilize fraud predictions across heterogeneous dealer, vehicle, and supplier datasets, improving score consistency under evolving data distributions.

- Built **large-scale risk stratification pipelines** on **Azure Databricks** using **PySpark** to continuously rank high-risk vehicles, dealers, and claims, outperforming legacy rule-based detection approaches in early-intervention accuracy.
- Developed **feature engineering frameworks** integrating vehicle telemetry aggregates, warranty histories, repair frequencies, dealer behavior signals, parts usage patterns, and geographic risk indicators, improving downstream model discrimination power.
- Applied **transformer-based NLP models** using **BERT**, **Hugging Face Transformers**, and **PyTorch** to analyze technician notes, warranty descriptions, and customer complaints, improving text classification accuracy by ~20% for investigative workflows.
- Designed scalable **NLP processing pipelines** with **spaCy** and **Azure AI Language** to perform entity extraction, root-cause categorization, sentiment analysis, and text normalization over unstructured service documentation.
- Implemented **document intelligence and OCR workflows** using **Azure Vision APIs** to extract structured data from scanned invoices, warranty forms, and inspection reports, significantly reducing manual review effort.
- Architected **secure, versioned ETL pipelines** using **Azure Data Factory** and **Delta Lake** to ingest, validate, and lineage-track warranty, dealer, and financial datasets supporting enterprise risk analytics.
- Leveraged **Azure AutoML** and **MLflow** to accelerate experimentation, hyperparameter tuning, and lifecycle tracking of fraud and risk models while maintaining reproducibility and audit readiness.
- Applied **explainable AI techniques** using **SHAP** and **LIME** to surface feature contributions and model rationale for high-impact fraud decisions, increasing investigator trust and supporting audit and compliance reviews.
- In later phases prototyped **LLM-assisted audit and investigation workflows** using **LangChain** and **Azure OpenAI** to summarize claim histories, surface policy references, and aggregate supporting evidence, reducing manual investigation effort by ~35%.
- Designed **batch and near–real-time inference workflows** to integrate ML risk scores into downstream warranty management, dealer oversight, and enterprise risk platforms with low operational latency.
- Containerized and deployed **ML inference services** using **Docker**, with exposure to **Azure Kubernetes Service (AKS)** and **CI/CD pipelines in Azure DevOps**, supporting reliable production rollouts and controlled rollback strategies.
- Implemented **model performance monitoring and basic drift detection** using **Azure Monitor** and scheduled evaluation jobs to ensure long-term stability and reliability of fraud and risk models.
- Developed **executive and operational dashboards** using **Power BI** to visualize fraud risk scores, model performance trends, dealer behavior patterns, and operational KPIs for leadership and business stakeholders.
- Collaborated closely with **warranty operations, finance, data engineering, and compliance teams** to productionize ML systems aligned with GM's enterprise risk controls, regulatory requirements, and real-world operational constraints.

Environment: Python, SQL, PySpark, XGBoost, Random Forest, PyTorch, TensorFlow, BERT, Hugging Face Transformers, spaCy, LangChain, Azure OpenAI, Azure ML, Azure Databricks, Delta Lake, Azure Data Factory, Azure Vision, Azure AI Language, MLflow, Docker, Azure Kubernetes Service (AKS), Azure DevOps, Power BI, SHAP, LIME

AI/ML Engineer **Jan 2019 – Dec 2020**

Smartous LLC, India

Enterprise Retail Analytics & Demand Forecasting Platforms

Responsibilities:

- Designed **ML-ready data pipelines** using **Python**, **PySpark**, and **AWS (S3, Glue, Redshift)** to support enterprise demand forecasting, pricing optimization, and inventory planning across thousands of SKUs and multiple fulfillment regions.
- Engineered **feature engineering frameworks** with **SQL**, **dbt**, and **Snowflake** to generate time-series, customer-level, and product-level features, improving downstream model stability across seasonal and high-volume product categories.
- Developed and deployed **demand forecasting systems** using **ARIMA**, **Prophet**, **LSTM**, and neural networks to model seasonality, trends, and promotional effects, improving forecast accuracy by ~25% for high-impact retail segments.
- Implemented **model evaluation and backtesting pipelines** using **RMSE**, **MAPE**, and rolling-window validation to compare statistical and deep-learning models prior to production rollout, reducing forecast error volatility.
- Built **ensemble forecasting approaches** combining statistical methods and neural models to improve robustness under demand spikes and promotional volatility, supporting more reliable inventory and replenishment decisions.
- Integrated **external business and market signals** including promotions, pricing changes, demand trends, and calendar effects as model features, improving sensitivity to short-term demand shifts and promotional accuracy.
- Operationalized forecasting models through **batch and scheduled inference pipelines** using **AWS Lambda** and **CI/CD workflows**, enabling reliable execution aligned with enterprise planning cadences.
- Designed **analytics and decision-support dashboards** using **Power BI**, **Tableau**, and **Amazon QuickSight** to surface forecast outputs, confidence intervals, and demand risk indicators for merchandising and supply-chain teams.
- Conducted **A/B testing, uplift analysis, and cohort analysis** to quantify the business impact of model-driven pricing and promotion strategies, informing data-backed rollout decisions.
- Implemented **data and model governance practices** using **AWS Lake Formation**, versioned datasets, and access controls to ensure reproducibility, auditability, and secure ML workflows.
- Collaborated with **DevOps teams** to productionize ML pipelines with **AWS CloudWatch** monitoring and alerting, improving pipeline reliability and operational visibility.
- Partnered with **merchandising, supply-chain, and e-commerce teams** to translate ML forecasts into actionable inventory, pricing, and replenishment decisions, increasing adoption of data-driven planning.

Environment: Python, SQL, PySpark, ARIMA, Prophet, LSTM, Neural Networks, AWS (S3, Glue, Redshift, Lambda, Athena, Lake Formation, CloudWatch), Snowflake, dbt, Power BI, Tableau, Amazon QuickSight, Git, CI/CD

ML Engineer **Sep 2017 – Dec 2018**

Cisco Systems, Chennai India

Predictive Analytics & Enterprise Risk Intelligence

Responsibilities:

- Designed **supervised predictive analytics systems** using **Python**, **scikit-learn**, **pandas**, **Random Forest**, and **Gradient Boosting** to support customer churn prediction, contract risk analysis, and anomaly detection across large enterprise customer portfolios.
- Built and evaluated **classification and ensemble models** including logistic regression, tree-based methods, and boosting techniques to identify high-risk customer accounts and abnormal usage patterns, reducing noise in downstream analytics and renewal workflows.
- Developed **feature engineering pipelines** using **SQL** and **Python** to construct customer-, contract-, usage-, and product-level features from CRM platforms, subscription systems, billing data, and network telemetry summaries.
- Implemented **customer segmentation and risk stratification analyses** using clustering algorithms and statistical techniques to surface churn drivers and usage behavior patterns, enabling more targeted renewal and retention strategies.
- Contributed to **automated ETL workflows** using **AWS Glue**, **Python**, and **SQL** to ingest, cleanse, and transform data from CRM, licensing, billing, and third-party systems into analytics-ready datasets.
- Supported **batch model training and scoring pipelines** using **AWS (S3, Redshift, Athena, Lambda, EC2)** to enable scalable risk scoring, periodic churn assessment, and enterprise reporting in a secure cloud environment.
- Optimized **analytical SQL queries** in **Amazon Redshift** to improve performance of recurring churn and risk reports, reducing query latency and improving dashboard responsiveness.
- Applied **statistical modeling techniques** using **SAS (PROC LOGISTIC, PROC REG)** alongside Python-based ML models to support pricing analysis, renewal likelihood estimation, and customer behavior insights.
- Built and maintained **executive and operational dashboards** using **Tableau** and **Power BI** to track churn trends, risk indicators, customer adoption metrics, and portfolio-level KPIs for analytics and business stakeholders.
- Participated in **model validation and monitoring activities**, tracking accuracy, stability metrics, and early drift indicators to support reliable production analytics and reporting outputs.
- Supported **enterprise data governance and compliance practices** by contributing documentation, data lineage, and control artifacts aligned with Cisco's internal risk, security, and analytics governance frameworks.
- Collaborated with **data engineers, business analysts, and DevOps teams** to integrate model outputs into reporting layers, renewal planning workflows, and decision-support systems.
- Gained hands-on experience across the **full ML lifecycle**, including data preparation, feature engineering, model development, evaluation, basic deployment, and monitoring within large-scale enterprise environments.

Environment: Python, SQL, SAS, scikit-learn, pandas, NumPy, Random Forest, Gradient Boosting, AWS (S3, Redshift, Athena, Glue, Lambda, EC2), Tableau, Power BI, ETL Pipelines

---

Python Developer/Data Analyst **Feb 2016 - Aug 2017**

Square Panda India Pvt Ltd

Telecom Analytics & Customer Intelligence

Responsibilities:

- Analyzed **large-scale telecom usage and operational datasets** using **Python (pandas, NumPy)** and **SQL** to support network performance monitoring, cost optimization, and customer intelligence across prepaid and postpaid segments.
- Designed **analytical data models and reporting datasets** integrating **call-detail records (CDRs)**, service logs, and customer profiles, improving consistency and reliability of downstream analytics used by operations and business teams.
- Built **customer behavior and churn analysis pipelines** using **Python** and **R**, identifying usage drop-offs, service-quality drivers, and lifecycle risk patterns to inform targeted retention strategies.
- Developed **predictive churn-risk and call-drop models** using **logistic regression** and **decision trees**, enabling proactive customer outreach and data-driven network optimization for high-risk cohorts.
- Engineered **automated ETL workflows** using **Python**, **SQL Server**, **MySQL**, and **Oracle** to ingest, cleanse, and standardize telecom data from multiple source systems, reducing manual preparation effort and reporting delays.
- Implemented **data validation, reconciliation, and anomaly-detection checks** within ETL pipelines to improve data accuracy, completeness, and trustworthiness for analytics and reporting workflows.
- Performed **market segmentation and cohort analysis** based on usage patterns, demographics, and service plans, supporting pricing optimization and customer acquisition initiatives.
- Prepared **labeled datasets and baseline features** for early churn prediction and network-quality models, supporting data science teams during experimentation and validation phases.
- Designed and executed **A/B testing frameworks** to evaluate service changes, feature rollouts, and customer-experience improvements, enabling statistically driven product and operational decisions.
- Conducted **VoIP and call-quality analytics** to identify network bottlenecks, dropped-call patterns, and service degradation issues, contributing insights for capacity planning and infrastructure optimization.
- Supported **database migration and consolidation initiatives** by validating data integrity, schema mappings, and historical consistency across legacy and modernized data platforms.
- Collaborated with **engineering, operations, and marketing teams** to translate analytical insights into actionable improvements across customer experience, network operations, and retention programs.

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

EDUCATION

Master's in computer science - University of Central Missouri | January 2021 – December 2022
Bachelor's in computer science - Lovely Professional University | August 2011 – May 2015

CERTIFICATIONS

- Microsoft Certified: Azure AI Engineer Associate - 2023
- AWS Certified Machine Learning - Specialty – 2021