

Prudhvi S Chundru

Lubbock, TX | (806)-559-6739 | prudhvyichundru01@gmail.com | [LinkedIn](#) | [GitHub](#)

PROFESSIONAL SUMMARY

Senior AI / Generative AI Engineer with 6+ years of experience designing and deploying large-scale ML and LLM systems across fintech and healthcare environments. Specialized in architecting **production-grade RAG platforms**, hybrid retrieval systems, and parameter-efficient fine-tuning workflows (**LoRA/PEFT**) supporting **multi-million transaction workloads**. Experienced in defining evaluation frameworks, reducing hallucination risk, and optimizing latency and token cost under strict compliance constraints. Strong background in **fraud modeling**, distributed data pipelines, and **cloud-native MLOps** across GCP and Azure. Focused on building reliable, scalable, and governance-ready AI systems that translate business objectives into measurable ML impact.

TECHNICAL SKILLS

- **Programming Languages:** Python, SQL, Java, Bash
- **Generative AI, LLMs & NLP:** GPT-4, Hugging Face Transformers, LangChain, Retrieval-Augmented Generation, FAISS, Pinecone, Qdrant, LLM Evaluation, LoRA/ PEFT (Pytorch), Cross-Encoder Reranking
- **Machine Learning & Deep Learning:** Scikit-learn, TensorFlow, PyTorch, XGBoost, LightGBM, Model Calibration
- **Data Engineering & ETL:** Apache Spark, PySpark, Airflow, Databricks, AWS Glue, Azure Data Factory
- **Backend & API Development:** FastAPI, REST APIs, WebSockets, gRPC
- **MLOps & Model Deployment:** MLflow, Docker, Kubernetes, AWS SageMaker, Azure ML, Google Vertex AI
- **Cloud Platforms:** Amazon Web Services, S3, RDS, Athena, Redshift, Google Cloud Platform, Vertex AI, BigQuery, Microsoft Azure, Azure OpenAI, Azure DevOps
- **Databases/Data Warehousing:** PostgreSQL, MongoDB, Amazon Redshift, Google BigQuery
- **Version Control & Collaboration:** Git, GitHub, Bitbucket, GitLab, JIRA, Confluence
- **DevOps & CI/CD:** Jenkins, Azure DevOps, GitLab CI, Terraform, Ansible, GitHub Actions

PROFESSIONAL EXPERIENCE

U.S Bank

Mar 2025 - Present

AI/ML & GenAI Engineer

- **Architected an enterprise RAG platform** indexing ~3TB of financial documents and supporting ~12M monthly transactions, using LangChain orchestration, FAISS/Pinecone vector indexing, and OpenAI GPT APIs for grounded response generation.
- Performed data exploration and preprocessing with Pandas and NumPy, engineered and selected features using Scikit-learn pipelines, which improved **model accuracy by 8%** and **cut training time by 15%**
- Evaluated embedding models (OpenAI vs open-source transformer embeddings) using precision@k, latency benchmarks, and cost-per-query analysis before production deployment.
- Implemented cross-encoder reranking using transformer-based models to improve **retrieval precision by 15%** and reduce hallucinated outputs.
- Conducted domain adaptation experiments using LoRA / PEFT fine-tuning in PyTorch, improving **classification accuracy by 9%** while **reducing GPU cost by ~60%** compared to full fine-tuning.
- Built a structured LLM evaluation framework measuring hallucination rate, grounding adherence, and latency. Implemented guardrails via retrieval confidence thresholds and post-generation validation filters.
- Redesigned fraud detection architecture processing ~5M daily transactions using XGBoost/LightGBM on distributed feature pipelines built with PySpark and BigQuery, improving **detection accuracy by 22%**.
- Introduced retrieval caching and token optimization strategies, reducing LLM API costs by ~28% without degrading answer quality.
- Designed CI/CD ML pipelines using **Vertex AI Pipelines, MLflow, and GitHub Actions**, reducing deployment cycle from weeks to <72 hours.
- Deployed containerized microservices using **Docker and Kubernetes (GKE)** with autoscaling and staged rollout (shadow → canary → full production), achieving **99.9% uptime** and reducing **incidents by 40%**.
- **Authored architectural design documentation** (ADR/RFC) for enterprise LLM platform, defining retrieval boundaries, scaling strategy, and governance model adopted across AI initiatives.
- **Mentored 3–5 engineers** on RAG design patterns, fine-tuning workflows, and evaluation methodology, establishing reusable internal GenAI standards.

Health Care Service Corporation

May 2024 - Mar 2025

AI/ML Engineer

- **Owned end-to-end architecture** of a **PHI-compliant RAG platform** indexing ~1.8M healthcare documents, defining embedding strategy, chunking policy, retrieval configuration, and evaluation benchmarks.
- Established LLM evaluation standards (precision@k, grounding accuracy, hallucination rate) and formalized acceptance criteria before production rollout.
- Led model selection analysis between Azure OpenAI and self-hosted transformer deployments, balancing latency, compliance, cost, and operational complexity.
- Designed and enforced data governance and redaction pipelines to ensure PHI-safe inference across LLM workflows.

- Defined model drift monitoring thresholds and retraining triggers using MLflow and Azure ML registry, reducing production **degradation incidents by 40%**.
- Coordinated integration with downstream care-management systems, ensuring low-latency inference within operational SLAs.
- Guided DevOps team in containerization and AKS deployment design, enabling controlled rollouts and rollback safety.
- Served as technical point of contact** for AI deployment decisions across analytics and infrastructure teams.
- Led cross-functional design reviews with compliance, infrastructure, and analytics teams to align LLM deployment with PHI governance and audit requirements.
- Established performance SLAs** (latency, grounding accuracy, and drift thresholds) and defined rollback criteria for safe production deployment.

TCS Global

ML Engineer

Jan 2021 - Dec 2023

Bangalore, India

- Led architectural redesign** of fraud detection pipelines processing **2+ TB of transactional data**, defining distributed Spark workflows and data partitioning strategies.
- Defined ensemble modeling framework (XGBoost, LightGBM) and threshold calibration strategy aligned with financial risk tolerance models.
- Established model evaluation and validation protocol incorporating cost-sensitive metrics and false-positive impact analysis.
- Designed scalable ingestion architecture using AWS Glue and Step Functions to ensure reliable multi-source data orchestration.
- Spearheaded migration of model training workflows to SageMaker distributed training, reducing **training time by 40%**.
- Owned deployment design** of Dockerized inference services integrated into enterprise decision engines.
- Partnered with risk and engineering stakeholders to translate business fraud objectives into measurable ML KPIs.
- Mentored junior engineers** on feature engineering and model optimization practices.

IBM

ML Engineer

Jan 2019 - Dec 2020

- Designed scalable ML pipeline architecture** using PySpark for enterprise anomaly detection use cases.
- Defined feature engineering and experimentation strategy improving **model accuracy by 12–15%**.
- Standardized ML experimentation framework** across projects, improving reproducibility and collaboration within engineering teams.
- Contributed to cross-team platform discussions to align ML pipeline design with enterprise cloud migration initiatives.
- Implemented monitoring and alerting standards using AWS CloudWatch to ensure production reliability.
- Automated infrastructure provisioning using Terraform, standardizing reproducible ML environments across projects.
- Collaborated with platform teams to integrate ML services into distributed enterprise systems.

PROJECTS

Intelligent Document Retrieval and Q&A System (RAG-based AI Assistant)

Jul 2025 - Oct 2025

Texas Tech University

Lubbock, TX

- Designed and implemented a **hybrid RAG architecture** combining dense embeddings (FAISS) and lexical search to improve multi-domain retrieval robustness.
- Benchmarked embedding models (OpenAI vs BGE-large vs E5) using precision@k and MRR to evaluate retrieval quality across varied query types.
- Implemented cross-encoder reranking to increase **precision@5 by 18%** without increasing retrieval depth (K).
- Conducted **LoRA fine-tuning experiments** in PyTorch (HF + PEFT) to adapt model behavior for domain-specific classification and formatting consistency.
- Built **hallucination mitigation layer** using retrieval confidence scoring and post-generation validation filters.
- Optimized token usage through dynamic chunk sizing and context window control, reducing **inference cost by 25%**.
- Deployed microservice-based inference pipeline using FastAPI + Docker + Kubernetes with async streaming responses.
- Implemented evaluation dashboard tracking latency, grounding rate, and response consistency.

Hope Speech Detection Project

Mar 2025 - Jun 2025

- Designed a multi-class NLP classification system on **~50K labeled social media samples** to detect hope speech under noisy, real-world text conditions.
- Addressed **severe class imbalance (~1:6 ratio)** using weighted cross-entropy and focal loss, improving **macro-F1** performance on minority class.
- Benchmarked traditional ML baselines (TF-IDF + Logistic Regression) against deep learning models (BiLSTM, BERT, RoBERTa), selecting transformer fine-tuning for optimal precision-recall balance.
- Fine-tuned transformer models in PyTorch using Hugging Face, optimizing learning rate schedules and batch size to **improve macro-F1 by 12%** over baseline.
- Performed stratified cross-validation and detailed error analysis to identify misclassification drivers such as sarcasm and contextual ambiguity.
- Calibrated classification thresholds to optimize precision-recall trade-offs depending on deployment scenario and exposed inference through a Dockerized FastAPI service.

EDUCATION

Texas Tech University

Master of Science (MS), Computer Science

Jan 2024 - Dec 2025

- **Coursework:** Software Testing and Quality Assurance, Data Structures and Algorithms, Automated Software Engineering, Machine Learning

Pragati Engineering College

Bachelors, Electronics and Communications

Jun 2016 - Dec 2020

CERTIFICATIONS

- **Solutions Architect Associate:** AWS
- **Meta Back-End Developer:** Coursera
- **Linux Fundamentals LFS101:** Linux Foundation
- **Python Programming Fundamentals:** Coursera
- **Test-Driven Development with Python:** Coursera
- **Python Automation with PyTest and Selenium:** Udemy
- **SQL for Data Science:** Udemy