Abhinav Sai Konjeti

LinkedIn | ☎ 716-259-4366 | Portfolio | abhi.konjeti@gmail.com | GitHub | Leetcode

Professional Summary

Senior Data Scientist and Generative AI Engineer with **7+ years of experience** in architecting and scaling LLM enabled, RAG, and agentic AI solutions that reduce operational costs and accelerate enterprise decision making. Demonstrated leadership in AI driven product innovation, LLMOps strategies, and multimodal AI, delivering measurable impact across SaaS, PaaS, and financial services platforms. Expertise spans LLM fine tuning, retrieval systems (Pinecone, Faiss, Chroma, LlamaIndex), inference optimization (Triton, vLLM, SageMaker, Vertex), and enterprise MLOps.

- **Enterprise Scale RAG & LLM Deployments**: Designed and optimized RAG frameworks with Pinecone/FAISS/Chroma and streamlined inference with Triton, vLLM, and SageMaker reducing latency by 40% and costs by 25%. Delivered multimodal assistants and domain specific LLMs that improved response relevance by 15% and enabled 92% accurate recommendations in loan servicing.
- **AI Powered Automation in Financial Services**: Built conversational AI and workflow assistants that eliminated 90% of manual claim handling, reduced service costs by 20–30%, and supported $50M+ loan approvals through AI driven credit risk insights and automated regulatory reporting compliant with FDIC/SEC standards.
- **Fraud & Compliance Innovation**: Engineered AI driven fraud detection and AML monitoring systems that boosted detection accuracy by 25%, cut false positives by 30%, and uncovered laundering patterns missed by traditional rules based approaches improving compliance reviews and operational efficiency.
- **Intelligent Data & Document Processing**: Delivered 95% accuracy in EMR extraction by combining OCR, HL7/FHIR parsing, and NLP pipelines. Accelerated BI & reporting by vectorizing 50M+ SQL warehouse records for natural language querying, reducing reporting cycles from weeks to minutes.
- **MLOps & Leadership**: Standardized ML deployments with Docker/Kubernetes and CI/CD pipelines for reproducibility. Migrated legacy platforms to AWS S3/EMR/Redshift, lowering costs by 15%. Mentored cross-functional teams on GenAI best practices, reducing project delivery time by 25% and accelerating enterprise adoption.

Technical Skills

Generative AI & Agentic AI :

- **LLMs:** GPT, Gemini, LLaMA, DeepSeek.
- RAG, Prompt Engineering, Model Fine-Tuning, LLM Inference Optimization (vLLM, Triton).
- **Agentic AI:** Multi Agent Systems, LangGraph, AutoGen, LangChain, LlamaIndex, LangSmith.
- AI Agent Security & Guardrails, Agent Communication Protocols, Agentic AI Architectures.

Machine Learning & Data Science :

- ML, DL, RL, NLP, Computer Vision, Feature Engineering, Model Deployment & Evaluation.
- **Frameworks:** TensorFlow, PyTorch, Scikit-Learn, Hugging Face Transformers, NLTK, OpenCV.
- **Data Analytics:** Pandas, NumPy, PySpark, Visualization (Matplotlib, Seaborn, Plotly).

Programming & Development :

- **Languages & APIs:** Python, REST APIs, GraphQL, OpenAI API.
- **Development Tools:** FastAPI, Streamlit, Neo4j.
- **Databases:** SQL, NoSQL, Vector Databases (Pinecone, FAISS, Chroma DB).

Cloud & MLOps:

- **Platforms:** AWS (SageMaker, Bedrock, Lambda, EC2), GCP (Vertex AI, BigQuery, Cloud Run), Azure (Azure ML, Synapse, App Services).
- **Practices:** MLOps/LLMOps, MLflow, AIOps, AutoML, Orchestration (Airflow, ADF, Dagster), CI/CD (Docker, Kubernetes, GitHub Actions).

Collaboration & Leadership:

- Agile/Scrum, Cross-Functional Team Leadership, Stakeholder Engagement.
- Business Intelligence, Product Strategy Alignment, Technical Writing & Documentation.

Experience

Generative AI & Machine Learning Engineer — UnitedHealthCare

Atlanta, GA — 03/2024 - present

- **AI Driven Healthcare MVPs & Enterprise Deployments:** Spearheaded MVP and production-grade initiatives, including a **Large Medical Model (LMM) for risk & cost prediction**, a **RAG powered BI insights pipeline**, and a **GPT-4o conversational AI for claims assistance**, directly reducing manual claim handling by **90%**.
- **End-to-End Document Intelligence Automation**: Designed OCR + NLP pipelines to extract structured data from **HL7, FHIR, faxes, and call transcripts**, automating intake for EMR and billing systems; achieved **95% accuracy in prior auth case adjudication**.
- **Scalable Cloud-Native AI/ML Solutions**: Built, deployed, and optimized AI pipelines on **AWS SageMaker, Bedrock, Lambda** and **GCP Vertex AI, BigQuery, Cloud Run**, enabling **real-time eligibility verification** and cost-efficient model serving.
- **Next-Gen Multi-Agent Architectures**: Engineered **agentic AI systems** with LangChain, LangGraph, Pinecone, and Graph DBs to autonomously handle eligibility checks, benefit verification, and compliance logic across insurance workflows.
- **Enterprise-Grade RAG Systems:** Implemented retrieval-augmented generation using **FAISS & vector stores**, enabling natural language querying of a **50M+ patient record SQL warehouse**, cutting reporting cycles from weeks to minutes.
- **Healthcare Analytics & BI Impact:** Delivered AI-powered dashboards for risk stratification and population health analytics, supporting **executive reporting** and improving provider decision-making on **50M+ covered lives**.
- **Responsible & Explainable AI:** Embedded bias testing, fairness audits, and **explainable AI** frameworks into the LMM lifecycle, ensuring compliance with Responsible AI principles and healthcare regulatory standards.

- **Cross-Functional Leadership**: Partnered with claims analysts, compliance officers, and engineers to align AI solutions with **business and regulatory goals**, improving productivity **4x** and delivering **10x ROI in OpEx savings**.
- **Innovation & Research Contributions**: Advanced **fine-tuning of LLMs** for healthcare vocabulary, evaluated **open-source and commercial models (GPT-4, Claude, Gemini)**, and contributed to Optum/UHC's research roadmap in **Generative AI & agentic systems**.
- **Key Achievements**: Delivered scalable enterprise AI systems that exceeded accuracy benchmarks, **accelerated deployment cycles by 40%**, and transformed **claims operations and member experience** through GenAI-driven automation.

Data Scientist & Generative AI Engineer (RAG Expert) — Frost Bank

San Antonio, TX 02/2023 – 03/2024

- **Implemented a Retrieval Augmented Generation (RAG) pipeline** integrating real-time Frost Bank data with a large language model, enabling context-aware customer-query responses. This RAG solution combined vectorized financial documents (loan policies, FAQs) and dynamic transaction data for instant retrieval improving GenAI productivity gains by 20%.
- **Engineered a "Virtual Loan Officer" chatbot** powered by RAG to automate end-to-end loan processing. This system used an LLM fine-tuned on Frost's financial corpus to answer customer questions, fetch required documents, and compute approval probabilities. In trials it delivered tailored financial guidance with ~92% accuracy in recommendations.
- **Cut average response time by 25%** in Frost's contact center by deploying AI-driven conversation summarization and real-time knowledge retrieval. This solution summarized customer conversations in real time and pulled answers from internal knowledge bases on demand.
- **Developed RAG-enhanced fraud-detection models** analyzing millions of transactions to flag anomalous patterns. These models automatically identified subtle signs of money laundering and fraud that traditional rule-based systems often miss.
- **Managed end-to-end data pipelines** using Python and Spark to ingest and preprocess large banking datasets (SQL/NoSQL, API feeds, customer transcripts). Key documents and records were embedded into a vector database (via semantic embeddings) to power fast similarity search. This allowed the RAG system to retrieve contextually relevant financial data within milliseconds.
- **Fine-tuned transformer-based models (BERT/GPT)** on AWS/GCP for classification and summarization tasks. Implemented a RAG workflow that continuously ingested the latest regulatory bulletins and policy documents via secure APIs, ensuring all AI outputs complied with current banking regulations. Model training and inference ran on cloud-native ML stacks (leveraging Python, PyTorch, TensorFlow) in accordance with Frost's technology standards.
- **Collaborated across cross functional teams** (IT, Risk, Customer Service) to align AI solutions with business goals. Translated complex model outputs (churn predictions, credit-risk scores) into clear insights for stakeholders. Regular presentations of data-driven recommendations helped drive strategy decisions in lending, marketing, and fraud prevention.
- **Maintained rigorous data governance** for all AI/ML processes. This included automated data cleaning, normalization, and validation routines to ensure high-quality inputs. Models were retrained on fresh data (new customer feedback, updated market info) every quarter, which kept performance stable and outputs accurate over time.
- **Piloted a self-service AI** assistant for routine banking tasks by connecting RAG to Frost's core APIs, automating low-value tasks like debit card replacement while freeing agents for complex cases resulted in 20–30% service cost reduction.
- **Adhered to Agile and MLOps best** practices throughout development. Model training and deployment were containerized (using Docker/Kubernetes) and integrated into CI/CD pipelines, reflecting Frost's emphasis on cloud-native AI/ML tooling. We also implemented monitoring and logging to ensure the RAG services met enterprise security and availability standards.

Data Scientist — AppMinds

HYD, India 08/2020 - 05/2022

- **Modernized legacy enterprise platforms** by designing AI-driven microservices on **AWS/Azure**, accelerating digital transformation and reducing infrastructure costs.
- **Engineered deep learning pipelines** using **PyTorch, CNNs, and ResNet** for image classification, anomaly detection, and document digitization across multiple client projects.
- **Built a FinTech KYC verification system** integrating OCR, face-matching CNNs, and fraud detection models, reducing customer onboarding time by **35%** and improving compliance accuracy.
- **Exposed REST API endpoints** for ML models (NLP, CV, recommendation engines), enabling seamless integration with mobile apps and enterprise platforms.
- **Developed Proof of Concepts (PoCs)** for AI-driven use cases (fraud detection, claims automation, predictive analytics), leading to client adoption of modernization roadmaps.
- **Constructed ETL pipelines and cloud data warehouses** (Snowflake, Azure Data Services) to unify heterogeneous datasets and support advanced analytics.
- **Designed RAG-powered chatbots** leveraging vector embeddings and domain-specific corpora, enhancing customer engagement and reducing manual service queries by **25%**.
- **Created KPI-driven dashboards in Power BI/Tableau** to track fraud-detection accuracy, onboarding efficiency, and customer retention, enabling data-driven executive decisions.
- **Optimized ML models** through hyperparameter tuning, pruning, and quantization, improving inference performance by **40%** in production.
- **Automated CI/CD pipelines** for ML models with **Docker and Kubernetes**, ensuring reliable deployments across multiple client engagements.
- **Delivered projects under the Build–Own–Transfer (BOT) model**, designing AI/ML solutions, managing early operations, and transferring full ownership to client teams.
- **Implemented NLP-based document processors** for text extraction, entity recognition, and compliance validation, reducing manual review workload by **30%**.
- **Mentored junior data scientists** on CNN architecture design, PyTorch best practices, and reproducible ML experiments to strengthen the internal talent pool.
- Led client workshops and solution demos**,** presenting PoCs and production-ready systems in fintech, healthcare, and enterprise domains, ensuring stakeholder buy-in and adoption.

Data Engineer **Yatra.com**

HYD, India 01/2018 - 07/2020

- Designed and **implemented ETL pipelines using Apache Spark and Scala** to process terabytes of flight, hotel, and booking data, reducing data processing times by **25%** and enabling timely analytics.
- Architected a **real-time streaming system with Apache Kafka and Spark Streaming** to deliver personalized travel offers based on user search and booking behavior, boosting engagement rates by **15%**.
- **Improved query performance by 20%** through developing distributed data warehouses with **Cassandra + HBase**, powering faster business reporting and analytics.
- **Cut query execution time by 30%** by optimizing **Hive SQL queries and pricing data models**, supporting faster pricing strategy decisions.
- **Enhanced travel search latency by 20%** by building and maintaining **Elasticsearch indices** for flights/hotels, improving overall user satisfaction.
- **Lowered deployment times by 10%** by managing Spark clusters on Databricks, streamlining large-scale pipeline execution and development.
- **Reduced operational overhead by 15%** by migrating legacy **Apache Storm workflows to Spark Streaming**, modernizing real-time data pipelines.
- **Enabled flexible customer segmentation** by storing semi-structured profiles in **MongoDB**, supporting marketing personalization campaigns.
- **Accelerated release cycles by 30%** with CI/CD pipelines (Jenkins, Git) and Dockerized deployments**,** ensuring consistent production rollout of data applications.
- **Strengthened fraud trend prediction & recommendations** by collaborating with data scientists to deploy ML models into production pipelines.
- **Increased reliability of downstream analytics by 20%** with **Airflow + Grafana monitoring**, automated validation, and proactive issue detection.
- **Reduced infrastructure costs by 15%** by migrating on-prem storage to **AWS S3, EMR, and Redshift**, improving scalability for peak travel seasons.
- **Improved Spark job performance & NoSQL queries** through extensive performance tuning, optimizing resource utilization on terabyte-scale datasets.
- Standardized knowledge sharing by documenting architectures, workflows, and procedures in **Confluence**, accelerating onboarding and collaboration.

Projects

- **WizDocAI(LLMs):** Developed a document intelligence platform using advanced NLP models such as BART and DistilBERT for text extraction, summarization, and sentiment analysis. Leveraged RAG and LLM integration to provide in-depth insights and interactive visualizations for PDF and DOCX files. Deployed via Streamlit and FastAPI, cutting document processing time by 60% and significantly improving efficiency in legal and academic workflows.
- **DocSummarizer (LLM):** Built an advanced document summarization system leveraging transformer-based LLMs such as Pegasus to generate concise summaries while retaining critical information. Integrated with FastAPI and Streamlit for real-time interaction, supporting PDF and DOCX formats, and employed semantic analysis for relevance extraction. Streamlined manual document review by 60%, enhancing productivity across legal and academic environments.
- **FMCG Daily Sales Prediction(ANN,PyTorch):** Developed a predictive sales model for FMCG products using a **feedforward neural network** using PyTorch. This dataset is taken from Kaggle, it consists of real world FMCG data of polish markets from 22-24. Applied MSE and Huber loss functions with dropout, early stopping, and ReduceLROnPlateau for robust training. Built an interactive Streamlit dashboard to visualize daily sales forecasts, improving inventory planning and demand prediction accuracy.
- **Industrial-Residential Air Quality Classification (ANN, PyTorch):** Developed a machine learning model to classify air quality levels in industrial and residential areas using sensor data. Implemented feature engineering techniques to enhance model accuracy. Used Binary crossentropy loss fucntion for this ANN. Deployed the model with Streamlit, providing real-time air quality monitoring and predictive analytics, aiding in environmental health assessments.
- **Customer Churn Prediction (ANN, PyTorch, Streamlit):** Developed an artificial neural network model to predict customer churn in the banking sector, utilizing features such as credit score, geography, age, and account balance. Employed BCE loss function, incorporating dropout layers and early stopping to enhance model generalization. Deployed an interactive Streamlit dashboard for real-time churn prediction, enabling proactive customer retention strategies.

Education

Master's in Data Science **University at Buffalo**

CERTIFICATIONS

- **Programming, Data Structures and Algorithms using python,** IIT Madras.
- **Data Mining**, IIT Kharagpur.
- **The Complete Data Structures and Algorithms Course in Python**, Udemy.
- **MySQL for Data Analytics and Business Intelligence**, Udemy.
- **Python basic-UOM**, Coursera.
- **Microsoft SQL Server**, NIT.
- **Python for Datascience**, IIT Madras.
- **Data Analytics with python**,IIT Roorkee.
- **Statistics for Data Analysis Using Python**, Udemy.
- **Introduction to Tensorflow for AI, ML, DL**, Coursera.
- **Python Programming**, NIT.