PRATIKSHA PAWAR

AI / ML Engineer

pratiksha@careeremail.net | (209) 734-3828 | USA | LinkedIn | GitHub | Portfolio

Eligible for Canadian work permit via Global Talent Stream (2-week processing) | Open to Toronto, Vancouver, Remote

SUMMARY

AI/ML Engineer with 4 plus years building and deploying production ML systems in financial services (Deutsche Bank) and healthcare (HCLTech). Expertise in RAG pipeline architecture, LLM fine-tuning, MCP-based agentic workflows, and end-to-end MLOps across AWS and Azure. Proven track record delivering LLM benchmarking, time-series risk forecasting, and compliance-grade model governance under Basel III and MiFID II. GitHub projects include a production-grade regulatory RAG pipeline and an MCP financial data assistant. MS in Data Science, SUNY Buffalo.

SKILLS

ML and Deep Learning: Python, R, Scikit-learn, XGBoost, LightGBM, TensorFlow, Keras, PyTorch, Pandas, NumPy, PySpark, ARIMA, Prophet, LSTM, SpaCy, TF-IDF

LLM GenAI and RAG: GPT-4, Gemini, Grok, Claude API, LLM benchmarking (TTFT, throughput, hallucination rate), LLM fine-tuning (LoRA, PEFT), Hugging Face Transformers, prompt engineering, RAG pipeline design, LangChain, LlamaIndex, LangGraph, ChromaDB, Pinecone, FAISS, embedding models, semantic chunking, hybrid BM25 plus vector retrieval

Agentic AI and MCP: Anthropic MCP, agentic workflow design, tool-use orchestration, multi-step reasoning, function calling, LangFuse, Weights and Biases, LLM observability and latency tracking

Data Engineering and ETL: SQL, NoSQL (MongoDB), Apache Spark, Airflow, AWS Glue, Informatica, Snowflake, Hadoop

Cloud and Deployment: AWS SageMaker, AWS EC2, AWS ECS, Azure ML, Azure Data Factory, Docker, Kubernetes, Flask, FastAPI, REST APIs, Microservices

Model Lifecycle and CI/CD: MLflow, Git, GitHub Actions, Jenkins, experiment tracking, model monitoring, drift detection, LLMOps

Compliance and Governance: HIPAA, GDPR, Basel III, MiFID II, OSFI, data anonymization, model governance, bias mitigation, audit logging

PROFESSIONAL EXPERIENCE

AI/ML Engineer, Deutsche Bank

10/2024 to Present | Remote, USA

- Engineered and deployed AI-driven market risk and trading models using Scikit-learn, XGBoost, and AWS SageMaker, improving portfolio return predictions by 26% and reducing trading execution latency by 15%
- Benchmarked GPT-4, Gemini, and Grok for production deployment on TTFT, throughput, hallucination rate, and cost; designed hybrid LLM architecture by risk level, reducing regulatory research time by 65%
- **Architected a RAG pipeline** over 10,000 plus regulatory documents (Basel III, MiFID II) using LlamaIndex and ChromaDB with semantic chunking and hybrid BM25 plus vector retrieval; reduced compliance analyst research time by 60% and achieved 91% relevance at top 5 on an internal evaluation set — code open-sourced on GitHub
- **Built an MCP server** integrating Claude with internal risk data tools, enabling analysts to run structured agentic queries across live trading and liquidity datasets via natural language, cutting time-to-insight from 15 minutes to under 90 seconds — architecture published on GitHub
- **Fine-tuned domain-specific LLMs** using LoRA and PEFT on Hugging Face Transformers for financial document classification, reducing hallucination rate by 34% versus base GPT-4 on regulatory Q and A tasks
- Automated trade and risk pipelines with Airflow and AWS Glue processing 25 million plus daily records, cutting manual ETL effort by 65% and runtime from 5 hours to under 50 minutes
- Designed time-series models (ARIMA, Prophet, LSTM) for market volatility and credit exposure forecasting, improving VaR and liquidity forecasting accuracy by 21%
- Containerized and deployed ML models using Docker and Kubernetes on AWS ECS with auto-scaling; implemented MLflow and LangFuse for lifecycle management, LLM observability, and regulatory-compliant version control
- Built CI/CD pipelines with Jenkins and GitHub Actions, shortening deployment cycles by 55%; documented model lineage and bias mitigation aligned with Basel III and MiFID II

AI/ML Engineer, HCLTech *01/2020 to 12/2023 | Pune, India*

- Built and deployed predictive models for patient readmission risk using Python and Scikit-learn, improving clinical decision accuracy by 20%
- Built automated data pipelines using Pandas, SQL, and Airflow to clean HIPAA-compliant EHR and claims data in Snowflake, achieving 98% data accuracy
- Built ETL workflows with Informatica and Spark integrating patient, claims, and lab data, reducing manual data preparation by 60%
- Applied SpaCy and TF-IDF NLP to detect ICD-10 coding anomalies, improving coding accuracy by 25%; deployed models as Flask microservices on AWS EC2, cutting deployment time by 75%
- Owned feature engineering and model evaluation for clinical predictive models; monitored production precision, recall, and drift metrics, maintaining 90% plus model consistency

GITHUB PROJECTS

Regulatory RAG Pipeline — github.com/prajendrapawar419/rag-regulatory-pipeline

- Hybrid BM25 plus vector retrieval system over Basel III and MiFID II regulatory documents using LlamaIndex and ChromaDB; achieves 91% relevance at top 5 on a 200-question compliance eval set — FastAPI served with citation-backed responses

MCP Financial Data Assistant — github.com/prajendrapawar419/mcp-financial-assistant

- MCP server exposing 6 financial tools (portfolio, risk metrics, trade history, liquidity ratios, positions, regional exposure) to Claude; enables natural language queries over live financial data — reduced analyst time-to-insight from 15 minutes to 90 seconds

EDUCATION

MS in Data Science, State University of New York at Buffalo *01/2024 to 05/2025*

BE in Electronics and Telecommunications, Savitribai Phule Pune University *08/2017 to 11/2020*

CERTIFICATIONS

- AWS Cloud Practitioner (Amazon Web Services)
- Certified Scrum Product Owner (CSPO) (Scrum Alliance)
- DeepLearning.AI - LangChain for LLM Application Development (Coursera)