BHARGAVI KADAMBARI

AI/ML Engineer

Irving,Tx,USA|469-5599646 | bhargavikadambari32@gmail.com| LinkedIn | Portfolio | GitHub |

Summary

AI/ML Engineer with 5+ years of hands-on experience building **end-to-end machine learning and GenAI systems** across document intelligence, recommendation engines, and **production-grade RAG pipelines.**
Expert in **Python, PyTorch, SQL, FastAPI,** and **cloud-deployed AI on AWS and GCP,** with deep practical experience in **LLMs, vector search (FAISS/Pinecone/ChromaDB), and multimodal AI pipelines.**
Proven impact in **reducing inference latency, improving contextual accuracy, and increasing user engagement** through retrieval optimization, personalization, and real-time AI workflows.

Technical Skills

- **AI / ML Core**: LLMs, Generative AI (GenAI), NLP, Computer Vision, RAG Architecture, AI Agents, Vector Embeddings, LLM Fine-Tuning (LoRA, QLoRA), Recommendation Systems, Time Series Forecasting, Supervised & Unsupervised ML, Statistical Analysis, A/B Testing, Causal Inference, Bayesian Methods
- **Deep Learning**: ANN, CNN, RNN, LSTM, Transformers, BERT, GANs, Multi-Head Attention, Few-Shot Learning, Neural Network Optimization
- **GenAI & Advanced LLM Systems**: LLM Deployment & Inference Optimization, Prompt Engineering (Few-Shot, Chain-of-Thought), Constitutional AI, Semantic Caching, KV-Cache Optimization, Contextual Memory & Personalization, Knowledge Graph Integration (Applied), Citation-Aware Grounding, Hallucination Detection, Prompt Injection Mitigation
- **Frameworks & Libraries**: PyTorch, TensorFlow, Keras, Scikit-Learn, Hugging Face Transformers, LangChain, FAISS, LLAMA, XGBoost, Model Context Protocol (MCP), OpenAI APIs, DeepSeek, Uma
- **MLOps & ML Engineering**: End-to-End ML & GenAI Pipelines, Model Training, Evaluation & Deployment, CI/CD for ML Systems, Docker, Kubernetes, Containerization, Model Monitoring, Drift Detection, Model Governance, Distributed ML Systems, Git, GitHub, Bitbucket, GitHub Actions
- **Cloud & Infrastructure**: AWS (S3, EC2, EKS, ECS, ECR, Lambda, Bedrock, SageMaker, EMR, Athena, Batch, Step Functions), Azure (Azure ML, AKS), Snowflake (integration-level), Microservices
- **Databases & Vector Stores**: MySQL, NoSQL Databases, Vector Databases (Pinecone, ChromaDB), FAISS
- **Programming**: Python, SQL, PySpark (ML-focused), R, Object-Oriented Programming (OOP), NumPy, Pandas
- **Multimodal & Real-Time AI**: Whisper STT, Multimodal Pipelines (Audio / Text / Limited Video), FFmpeg, Streaming-Based LLM Pipelines
- **Monitoring & Observability**: Prometheus, Grafana, OpenTelemetry, Sentry, Latency Monitoring, Error Tracking, Model Drift Detection, Prompt Safety Monitoring
- **AI Safety, Privacy & Compliance**: PII Masking, Output Verification, Rate Limiting, AI Safety Controls, Trust & Reliability in GenAI Systems
- **Operating Systems**: Linux, UNIX,Ubuntu, Windows, macOS
- **Developer & Productivity Tools**: Cursor, Windsurf, IntelliJ IDEA, Visual Studio, GitHub Copilot

Work Experience

Invent Artificial LLC **2024 - 2025**

GenAI Engineer Intern *Texas*

- Reduced **average LLM inference latency by ~40%** by optimizing RAG execution with **semantic caching, KV-cache reuse, async batching, and embedding quantization**, significantly improving end-user response times in interactive AI workflows
- Improved **task completion and contextual response accuracy by ~30%** by enhancing **retrieval-augmented generation (RAG)** pipelines with **hybrid dense retrieval, metadata-aware chunking, and reranking** using **Pinecone and ChromaDB**.
- Increased **user engagement by ~30%** by implementing **personalization and conversational memory layers**, combining session context, behavioral signals, and **vector similarity ranking (FAISS)** to deliver context-aware AI responses.
- Built and optimized **FAISS-based vector retrieval layers** for large unstructured document corpora, enabling **low-latency semantic search and citation-aware contextual grounding** for GenAI applications.
- Engineered **real-time multimodal AI pipelines** (STT → LLM → TTS) using **Whisper, GPT models, and streaming APIs**, improving conversational coherence and responsiveness during live interactions.
- Engineered **real-time media processing pipelines** using **Kafka, Flink, FFmpeg, and Whisper STT** to support **multimodal summarization (audio/text and limited video streams)**, enabling low-latency ingestion, transcription, and downstream LLM-based analysis.
- Implemented **behavior- and gesture-based fraud detection algorithms** by analyzing **interaction patterns, response timing, and non-verbal signals**, applying anomaly detection to flag suspicious behavior in AI-driven interview environments.
- Automated **GenAI deployment and MLOps workflows** using **AWS SageMaker, Docker, and GitHub Actions**, integrating CI/CD pipelines, model versioning, and evaluation workflows for iterative LLM releases.

- Integrated **observability and monitoring for GenAI services** using **Prometheus, Grafana, OpenTelemetry, and Sentry** to track latency, errors, model drift, and prompt-injection patterns, improving incident detection and reducing issue resolution time by **~45%**.
- Integrated **AI safety, compliance, and privacy controls**, including **hallucination detection, PII masking, prompt-level filtering, rate limiting, and output verification**, strengthening trust and reliability of GenAI systems in production.

Wipro Limited **2021 - 2023**
AI/ML Engineer *India*

- Analyzed **large-scale website interaction and clickstream data** using **SQL and Python**, uncovering deep navigation and engagement patterns and reducing manual exploratory analysis effort by **~30%** through reusable analytics workflows.
- Engineered **robust data preprocessing and feature engineering pipelines** on behavioral datasets (session duration, page sequences, interaction frequency), improving **ML feature quality and model stability by ~20%**.
- Built and evaluated **supervised ML models** including **Logistic Regression, Random Forest, and Gradient-based classifiers** to predict user engagement propensity, improving **engagement prediction accuracy by ~15%**.
- Applied **unsupervised learning techniques** such as **K-Means clustering** to segment users based on browsing and interaction behavior, enabling targeted content strategies and improving **personalization precision by ~15%**.
- Designed and implemented a **content-based recommendation system** leveraging **feature similarity and behavioral signals** to recommend relevant pages, reports, and documents, increasing **content discoverability and user engagement by ~12%**.
- Performed **model benchmarking, validation, and bias analysis** using **Decision Trees and rule-based baselines**, reducing low-confidence or noisy recommendations by **~10%** and improving overall recommendation reliability.
- Developed **Flask-based RESTful APIs** to serve real-time ML predictions and recommendation outputs, enabling seamless integration with the website platform and reducing **manual reporting and integration effort by ~25%**.
- Built **interactive analytical dashboards** using **Tableau** to visualize user segments, engagement funnels, and recommendation performance metrics, accelerating data-driven decision-making and reducing reliance on intuition by **~20%**.

TCS ion **2020 - 2021**
ML Engineer Intern *India*

- Built an **end-to-end ML-driven document digitization pipeline** combining **Computer Vision, OCR, and NLP**, automating extraction from scanned PDFs and images and reducing **manual document review effort by ~35%** across internal workflows
- Applied **image preprocessing and enhancement techniques** using **OpenCV** (noise removal, thresholding, skew correction) to improve **OCR text extraction accuracy by ~20%**, significantly reducing downstream validation and correction errors
- Implemented **BERT-based text classification and information extraction models** to automatically categorize documents and extract key fields, improving **document routing and classification accuracy by ~25%**.
- Developed **ML-based validation and consistency-checking models** to cross-verify extracted entities against document context and business rules, reducing reprocessing cycles and shortening **document processing turnaround time by ~30%**.
- Designed and deployed the AI pipeline as a **Flask-based RESTful service**, enabling seamless integration with internal applications and improving **end-to-end document processing efficiency by ~15%**.

PROJECTS

SMART RAG – Educational AI Assistant

- Built an educational RAG assistant that ingests PDFs/web content and delivers source-grounded answers using semantic search, vector databases, and OpenAI-powered generation via FastAPI. **GITHUB**.

FinGuard – Clean Fraud Detection API Platform

- Built a production-ready fraud detection API using Python, scikit-learn, and FastAPI, combining ML models with rule-based risk logic to deliver clear, real-time transaction risk assessments. GITHUB

Customer Churn Prediction MLOps Pipeline – Telco Churn API

- Built a full MLOps churn prediction pipeline with MLflow-tracked Random Forest models, a FastAPI inference service, and production-style monitoring using Prometheus, Grafana, Evidently, Docker, and Kubernetes. GITHUB

Interview Cheating Detection API – Real-Time Monitoring System

- Built a real-time Interview Cheating Detection API using FastAPI and Cursor AI, translating high-level requirements into detectors for screen, browser, input, process, and network activity through structured prompt engineering. GITHUB

Stock Price Forecasting – Multi-Model

- Built a multi-model stock price forecasting pipeline using ARIMA/auto-ARIMA with rich visual analytics to predict market trends and compare forecast performance on large OHLCV datasets. GITHUB

NYC 311 End-to-End Data Pipeline

- Built a GCP data pipeline using GCS ’Dataproc PySpark ETL ’BigQuery analytics, ending with an ARIMA_PLUS forecasting model for NYC 311 service request volumes. GITHUB

CERTIFICATIONS

- Microsoft Certified: Azure AI Fundamentals (AI-900)
- Google Cloud Certified Professional Data Engineer
- Oracle Database SQL Certified Associate (1Z0-071)

EDUCATION

University Of North Texas **2024 - 2025**
Masters, Data Science, Applied Natural Language Processing & Generative AI

Jawaharlal Nehru Technological University **2017 - 2021**
Bachelors, Electronics & Communication