Bhargavi Kadambari
TX, USA | 469-351-9574 | bhargavikadambari32@gmail.com | LinkedIn

SUMMARY:

AI/ML Engineer with 4+ years of experience in cloud data engineering and advanced machine learning solutions. Strong expertise in LLMs, NLP, GenAI, deep learning, and RAG architectures, with hands-on work in LangChain, vector databases, knowledge graphs, and MLOps. Skilled at building real-time AI systems for fraud detection, interview agents, and predictive analytics using AWS & GCP.

SKILLS:

- **AI/ML Core:** LLMs, NLP, Computer Vision, GenAI, RAG Architecture, AI Agents, Vector Embeddings, LLM Fine-Tuning, Causal Inference, Bayesian Methods, Statistical Analysis, A/B Testing, ML Algorithms, Time Series Forecasting, Recommendation Systems.
- **Frameworks & Tools:** PyTorch, TensorFlow, Keras, Scikit-Learn, HuggingFace Transformers, LangChain, LLAMA, XGBoost, Model Context Protocol (MCP), OpenAI, DeepSeek, Uma.
- **Deep Learning:** ANN, CNN, RNN, LSTM, Transformers, BERT, LoRA, QLoRA, Multi-Head Attention Mechanisms, Neural Network Optimization, Few-Shot Learning, GANs.
- **Data Science & Analytics:** EDA, Data Cleaning, Preprocessing, Feature Engineering, Statistics, Matplotlib, Seaborn, Excel Charts, Tableau, Looker Studio.
- **GenAI & NLP Advanced:** LLM Deployment, Prompt Engineering, Chain-of-Thought, Constitutional AI, Semantic Caching, Knowledge Graphs.
- **MLOps & Engineering:** End-to-End ML Pipelines, Model Training, Testing & Deployment, CI/CD, Docker, Kubernetes, Containerization, Model Monitoring, Model Governance, Git, GitHub, GitHub Copilot, Bitbucket, Distributed Computing.
- **Cloud & Infrastructure:** AWS (S3, EC2, EKS, ECS, ECR, Lambda, Bedrock, SageMaker, EMR, Athena, Batch, Step Functions, Redis), Azure (AML, AKS), Snowflake, Microservices.
- **Data Engineering:** ETL Pipelines, Big Data Processing, Hadoop, Spark, Kafka, Apache Airflow, Real-Time Systems, Data Governance.
- **Databases:** MySQL, NoSQL, Vector Databases (Pinecone, ChromaDB)
- **Programming:** Python, SQL, PySpark, R, OOP Concepts, Numpy, Pandas
- **Developer & Productivity Tools:** Cursor, Windsurf, IntelliJ, Visual Studio, GitHub Copilot.

WORK EXPERIENCE

Invent Artificial LLC – AI/ML Engineer Intern **Jan 2025 – Present**

- Architected a Retrieval-Augmented Generation (RAG) pipeline within the AI Interview Platform, using vector databases (Pinecone, ChromaDB) for semantic retrieval, achieving 92% contextual accuracy in candidate Q&A.

- Integrated Model Context Protocol (MCP) to enable pluggable tools (NLP, knowledge graphs, evaluation modules), making the platform extensible across multiple domains.
- Engineered a multi-modal pipeline with speech-to-text, sentiment analysis, and LLM-driven adaptive questioning, improving candidate engagement and response relevance by 40%.
- Fine-tuned LLMs (OpenAI, DeepSeek, Uma) with LoRA/QLoRA adapters and implemented constitutional AI guardrails, ensuring ethical, bias-reduced, and domain-aligned interview interactions.
- Built a knowledge graph reasoning engine that linked candidate responses to domain-specific concepts, enabling personalized follow-up questions tailored to technical and behavioral skills.
- Optimized the platform's inference pipeline with semantic caching, hierarchical transformers, and multi-head attention, reducing interview processing latency by 60%.
- Developed evaluation and feedback ML models to score candidate clarity, correctness, and confidence, visualized through Seaborn dashboards for recruiters.
- Deployed the platform on AWS (EKS, ECS, Fargate, SageMaker, Lambda, and S3) using Dockerized microservices and CI/CD pipelines, enabling scalable and production-grade delivery.
- Designed an AI Interview Agent with TTS integration, acting as a virtual interviewer that could conduct conversations and deliver real-time feedback summaries.

Data Engineer – GCP | Wipro Technologies **2022 – 2023**

- Designed and implemented ETL pipelines using Dataflow (Apache Beam) and Dataproc (PySpark) for large-scale batch and streaming workloads.
- Migrated legacy systems into BigQuery, applying partitioning/clustering strategies that reduced query latency by 30% and improved cost efficiency.
- Built real-time streaming pipelines with Pub/Sub + Dataflow, supporting predictive analytics and anomaly detection with near-zero latency.
- Automated workflow orchestration with Cloud Composer (Airflow), ensuring reliability in data delivery and ML-ready feature pipelines.
- Developed feature-engineering datasets in collaboration with data scientists, accelerating ML model training and deployment.
- Deployed data quality checks and governance using GCP Data Catalog and custom validation scripts.
- Optimized Spark-based transformations for high-volume datasets, improving overall pipeline performance.
- Delivered cloud-native, scalable architecture that improved reporting agility and supported AI/ML experimentation across business domains.

Data Engineer – AWS | Wipro Technologies **2021 – 2022**

- Architected a cloud-based data lake on Amazon S3, centralizing raw and processed data for analytics and ML workloads.
- Designed and automated ETL pipelines using AWS Glue + Step Functions, cutting processing times by 45%.
- Built real-time ingestion pipelines with Kafka + AWS Lambda, reducing reporting latency by 40%.
- Integrated Amazon Redshift for optimized transformations and incremental loading, improving query performance by 35%.

- Orchestrated workflows with Apache Airflow, ensuring scalable, production-ready reporting pipelines.
- Partnered with data science teams to prepare feature-engineering datasets for ML model training and deployment.
- Implemented pipeline monitoring, validation, and governance to ensure reliability and compliance.
- Delivered end-to-end cloud-native architecture that accelerated predictive analytics and real-time reporting.

Application Engineer – TCS iON **2020 – 2021**

Responsibilities:

- Developed and deployed containerized Java applications using Docker, ensuring portability and consistency across development, testing, and production environments.
- Configured and maintained Docker Compose for orchestrating multi-container applications, streamlining integration of databases, APIs, and services.
- Designed and implemented microservices-based architecture**,** improving scalability, fault isolation, and modular development.
- Automated build, testing, and deployment pipelines using CI/CD workflows, reducing manual effort and cutting deployment time by 40%.
- Integrated Git-based version control and container registries to manage and distribute container images.
- Conducted performance optimization and container lifecycle management, improving system efficiency and resource utilization.
- Built cloud-ready deployment environments as a foundation for future AI/ML pipelines and data engineering workloads**.**
- Collaborated with cross-functional teams to apply DevOps best practices, ensuring reliability and faster release cycles.

ACADEMIC PROJECTS

FinGuard – AI Fraud Detection Platform

- Designed and deployed a fraud detection system using ensemble ML models (Logistic Regression, Random Forest, Neural Networks), achieving 99% detection accuracy on high-volume transaction data
- Implemented Explainable AI (XAI) techniques with SHAP and LIME for transparent fraud scoring and regulatory compliance
- Built real-time inference pipelines (<10ms latency) with FastAPI, REST APIs, Docker, and CI/CD workflows, ensuring scalability and high availability in production
- Integrated data preprocessing, feature engineering, and anomaly detection techniques for improved model performance
- Leveraged cloud deployment strategies with AWS Lambda, ECS, and SageMaker, supporting automated retraining and monitoring.

NYC 311 Data Pipeline & Forecasting:

- Designed an end-to-end cloud data pipeline on Google Cloud Platform (BigQuery, Dataflow, Dataproc, Cloud Composer) to process millions of NYC 311 service requests
- Automated ETL workflows using PySpark on Dataproc, handling structured and unstructured datasets with data validation and governance checks
- Built a centralized data warehouse on BigQuery, optimizing queries with clustering/partitioning for cost efficiency and faster insights
- Applied time-series forecasting models (ARIMA, LSTM, Prophet) to predict request volumes, enabling proactive resource allocation for city services
- Developed dashboards in Looker Studio and Seaborn to visualize service demand trends and predictive insights.

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** **Jan 2024 - Present**
  Masters in Data Science

- **Jawaharlal Nehru Technological University** **July 2017 - Jun 2021**
  Bachelors in Electronics & Communication