KAVYA BALTHA

☎ 571-668-3223 ✉ kdlskavya@gmail.com in KavyaBaltha kavya507 Portfolio

OBJECTIVE

Generative AI and ML specialist with expertise in LLM applications, fine-tuning, and RAG-based semantic search. Proficient in building production-grade AI systems using FastAPI, LangChain, AI agents, and vector databases.

SKILLS

Languages & Frameworks	Python, SQL, JavaScript, PyTorch, TensorFlow, Hugging Face, FastAPI, Spark
Tools	Docker, Kubernetes, MLflow, Weights & Biases, GitHub, Airflow
GenAI & MLOps:	LLaMA 2, OpenAI, QLoRA, Chroma, RAG, LangChain, Google ADK

WORK EXPERIENCE

Data Scientist Aug 2023 - Present

The Cigna group

- Developed a Proof-of-Concept (POC) for an AI-powered search using Retrieval-Augmented Generation (RAG). Integrated semantic embeddings extracted from SQL databases and managed with Chroma vector database to deliver precise and contextually relevant information.
- Engineered XGBoost models to accurately predict patient readmissions and hospitalization durations, leading to a 20% improvement in patient segmentation.
- Automated ingestion pipelines built using AWS Glue and Athena, decreasing data processing latency by 30%. Visualized clinical KPIs using Tableau to guide leadership decisions.
- Conducted A/B testing for member engagement strategies, analyzing the effectiveness of digital outreach to identify optimal communication methods, leading to increased member interaction and retention.

Senior Technical Associate June 2019 - Aug 2021

Bank of America *Hyderabad, India*

- Designed and optimized data models to manage and analyze large volumes of trade data, utilizing machine learning techniques, including regression analysis, to enhance data retrieval performance by 20%.
- Analyzed trends in high-frequency trade data, employing predictive analytics to support risk management and operational efficiency, leading to a 15% reduction in operational costs.

PROJECTS

LLM Review Summarizer Chrome Extension

- Built a FastAPI-based backend paired with a Chrome extension to process user-supplied reviews and generate structured pros and cons summaries.
- Designed and deployed a fully agentic system using Google ADK, where a Tool-Selecting Agent orchestrates LLM calls for translation, sentiment classification, Summarization using GPT-4o. (Try it here)

Fine-Tuning LLaMA2 with QLoRA (4-bit)

- Fine-tuned LLaMA-2-7B-chat using QLoRA (4-bit quantization with LoRA, r=16) on the MeQSum dataset, significantly enhancing summarization accuracy and efficiency.
- Leveraged Hugging Face, PEFT to optimize training efficiency, achieving a 60% reduction in GPU memory usage and a 40% reduction in training time without compromising summarization accuracy. (Medium Blog)

EDUCATION

Master's in Computer Science, George Mason University, Virginia, USA Aug 2021 - May 2023

Bachelor's in Computer Science and Engineering, JNTUH, India June 2015 - May 2019