Muhammad Ali Azeem

m.ali.azeem84@gmail.com | (415) 818-4132 | Pleasanton, CA | https://www.linkedin.com/in/muhammad-aliazeem/

EDUCATION

Bachelors In Architecture 2003 - 2007
National College Of Arts, Punjab, Pakistan

PROFESSIONAL EXPERIENCE

Senior AI/ML Engineer Dec 2023 - Present
Workday, Pleasanton, California

- Designed and implemented a Retrieval-Augmented Generation (RAG) pipeline to reduce customer support call time by enabling intelligent, context-aware responses.
- Implemented document segmentation and dense vector embeddings using sentence transformers, enabling efficient semantic search and optimized knowledge retrieval in large-scale text corpora.
- Implemented ChromaDB for managing and indexing vector embeddings, and later migrated to MilvusDB to achieve greater robustness, scalability, and fault tolerance in handling large volumes of high-dimensional data with low-latency access.
- Designed and implemented a hybrid information retrieval system that integrated traditional keyword-based methods with modern embedding-based approaches, resulting in improved accuracy, relevance, and robustness across a wide range of queries.
- Applied prompt engineering techniques to optimize response generation, enhancing accuracy, consistency, and overall quality of outputs aligned with objectives and user requirements.
- Designed and deployed agentic AI workflows to automate customer support tasks
- Implemented Model Context Protocol (MCP) to seamlessly integrate LLMs with enterprise data sources, APIs, and tools, enabling real-time contextual awareness
- Integrated a orchestration agent to dynamically route tasks to the most suitable specialized agent, improving efficiency, reducing errors, and enabling seamless multi-agent collaboration
- Integrated task orchestration with LangGraph and FastAPI, enabling dynamic agent collaboration and delivering scalable, low-latency conversational experiences
- Designed and executed an evaluation framework using Ragas, retrieval, generation, and agentic workflow assessments to measure accuracy, reliability, and real-world performance with human-in-the-loop feedback
- Implemented NLTK-based text classification and summarization to establish baseline intent recognition and accelerate customer query resolution.
- Applied named entity recognition (NER) techniques to extract key information from customer queries, improving context understanding and downstream classification accuracy.
- Fine-tuned BERT LLM for classification tasks using Hugging Face Transformers to improve model performance and domain adaptation
- Leveraged Databricks to fine-tune BERT LLMs, utilizing Spark-based distributed computing to efficiently process large-scale datasets and optimize training workflows
- Implemented end-to-end deployment pipelines on SageMaker to serve ML models at scale with robust monitoring and auto-scaling
- Developed scalable backend services with FastAPI, integrated Redis for caching and DynamoDB for chat history storage, ensuring low-latency conversational experiences.
- Containerized and deployed applications on AWS EKS using Docker and Kubernetes, leveraging Helm charts for configuration management and GitHub-based CI/CD pipelines to ensure scalable, automated, and reliable orchestration of production workloads

Senior Software Engineer / Mlops Sept 2021 - Oct 2023
Lending Club, San Francisco, California

- Implemented ML pipelines for clustering and segmentation of large-scale datasets, leveraging AWS for efficient querying and integration with ETL workflows.
- Collaborated with data scientists to productionize unsupervised learning models for enterprise deployment.
- Leveraged MLflow for experiment tracking, model versioning, and reproducibility across multi-environment workflows.
- Implemented DVC-based pipelines to version datasets, track data lineage, and ensure reproducibility across multiple environments.
- Managed data storage and retrieval in AWS S3, ensuring efficient handling of large datasets and seamless integration with ETL processes.

- Created modular feature engineering workflows for numerical, categorical, and behavioral datasets.
- Enabled real-time monitoring and drift detection to ensure consistent model performance in production.
- Ensured data quality, consistency, and integrity across multiple sources, improving downstream ML reliability.
- Containerized ML models using Docker for consistent deployment across staging and production environments.
- Deployed containerized ML models using FastAPI, enabling consistent deployment across environments and providing real-time inference for downstream applications.

Software Engineer May 2019 - Sep 2021

Rally Health, San Francisco, California

- Built patient-facing and provider-facing healthcare applications using React.js, Node.js, and TypeScript to improve accessibility and usability
- Designed and deployed cloud-native applications on Google Cloud Platform (GCP), utilizing Cloud Run, Pub/Sub, and Cloud Functions
- Worked on secure health data pipelines (HIPAA compliant), ensuring sensitive data integrity with GCP encryption and IAM policies
- Implemented automated testing frameworks (Cypress, Jest, Mocha) to improve code quality and reduce release issues
- Modernized legacy monolith services into microservices deployed with Kubernetes (GKE)
- Designed real-time healthcare event tracking using Pub/Sub and Cloud Functions, enabling faster alerts and notifications
- Creation of Docker Images with the features which are required to run any browser tests and push it to Artifactory with the latest version tag.
- Write code in groovy and create global functions which can be used in Jenkins file for the Jenkins instance
- Develop and maintain multiple branch pipelines for each project for CI/CD Jenkins.
- Collaborated with cross-functional teams in Agile/Scrum environment, delivering high-quality and efficient releases.
- Worked closely with product managers and stakeholders to refine features supporting patient well-being.

Sofware Engineer June 2015 - May 2019

Autodesk, San Francisco, California

- Developed scalable web applications using AngularJS, React.js, and JavaScript ES6.
- Built RESTful APIs and backend services using Node.js, Express.js, and MongoDB
- Implemented microservices architecture for modular, scalable, and maintainable applications
- Optimized SQL and NoSQL queries to improve application performance and scalability.
- Integrated applications with Google Cloud Platform (GCP) services such as Cloud Storage, Pub/Sub, and BigQuery
- Applied design patterns (Observer, Singleton, Factory) to enhance code reusability and maintainability.
- Enhanced application reliability through unit testing (Mocha, Jasmine, Karma) and conducting code reviews
- Developed interactive dashboards using Google Charts and Chart.js to visualize key business metrics and trends

SKILLS

Languages	Python, Java, JavaScript
Frameworks	Scikit-learn, PyTorch, MLflow
Libraries	Langchain, Llamaindex, LangGraph, NLTK, Transformers, Scikit-learn, NumPy, Pandas, HuggingFace
LLMs	Anthropic Claude, OpenAI, BERT, LLaMA, BART
Database	SQL, PostgreSQL, MongoDB,Redis, dynamoDB
AWS	EKS, Bedrock, RDS, EC2, CloudWatch, Lambda, S3, Sagemaker, QuickSight
VectorDB	ChromaDB, PgVector, MilvusDB, FAISS
CICD	Jenkins, Docker, Kubernetes, GitHub Actions