Adarsh Balanolla

(469)943-6644 | adarshreddybms@gmail.com | LinkedIn | GitHub

EXPERIENCE

AI/ML Engineer | Tessolve Semiconductors, San Jose, CA **Feb 2025 – Present**

- Leading revolutionary GenAI-assisted EDA (Electronic Design Automation) platform development using LLMs, custom AI Agents and MCPs, targeting 60% reduction in PCB design cycles through natural language interface.
- Created BOM-to-schematic generator with internal component library, KiCad and LCSC API, simplifying complex workflows from weeks to minutes for Silicon Validation engineers.
- Deployed on-premises GenAI platform using vLLM for local inferencing, Docker for orchestration, and FastAPI, empowering engineers to overcome security constraints, while processing 5,000+ requests/day at 99.99% uptime.
- Designed an AI Agent to optimize PCB component routing, leveraging heuristic algorithms and ML-driven pathfinding to enhance layout efficiency and reduce design cycle time.
- Integrated the GenAI system with internal EDA (Electronic Design Automation) tools, enabling seamless data exchange and improving real-time assistance in design verification (DRC) workflows.

AI/ML Engineer | Flexon Technologies, Pleasonton, CA **Jul 2024 – Feb 2025**

- Developed RAG-based multimodal GenAI application leveraging LLMs (GPT-4o), Kore.ai, NLP, and Vector Database (Pinecone) to enhance IVR and speech-to-text capabilities for improved customer experience.
- Designed system prompts, implemented GenAI testing with the RAGAS framework, and performed human evaluation to enhance performance and user alignment.
- Implemented vector similarity search using Pinecone for real-time recommendation and hybrid (semantic + keyword) search capabilities, enhancing the relevance of user interactions.
- Engineered end-to-end MLOps pipelines integrated CI/CD workflows using AWS CodePipeline, Elastic Kubernetes Service (EKS), and ECR to automate deployment and version control, achieving 99.9% uptime.
- Developed and deployed RESTful APIs with FastAPI for interaction with GenAI models, enabling real-time query processing and front-end integration while improving response times and scalability.

Data Engineer | Flexon Technologies, Remote. **Sep 2023 – Jun 2024**

- Optimized conversational chatbot designs in Kore.ai, configuring intents, entities, and dialogue flows to enhance user experience across voice and chat platforms.
- Collaborated with DevOps teams to ensure seamless integration of CI/CD pipelines with Kubernetes and Docker, optimizing infrastructure for real-time applications.
- Improved response quality by 30% using prompt engineering techniques, i.e., Few-Shot, Chain-of-Thought (CoT), context window optimization, and instruction refinement to reduce hallucinations.

Data Engineer | Unosis, Bengluru, IND **May 2019 – Aug 2021**

- Automated data pipeline workflows using AWS Glue, PySpark, and SageMaker-Data Wrangler, reducing data preparation time and improving data freshness for real-time analytics.
- Engineered and implemented an AI-powered chat/voice bot using Amazon Lex to replace the legacy IVR system, automating customer support processes and leading to a 60% reduction in manual intervention.
- Optimized ETL pipelines using AWS services (S3, Glue, Lambda) and Snowflake to extract data from Salesforce email campaign operational data, ensuring automated T-1 data extraction, transformation, and loading of data.
- Developed Power BI dashboards with Snowflake & Star Schema models, reducing data redundancy by 24%, and engineered a data model with Microsoft SQL Server, SSRS, and SSIS, increasing customer retention.

SKILLS

Languages: Python (Pandas, PySpark, Scikit-learn, TensorFlow-Keras, NLTK), SQL, CQL, Cypher, R.
Databases: Pinecone, ElasticSearch, Microsoft SQL Server, PostgreSQL, Supabase, MongoDB, Neo4j, Cassandra
Machine Learning & Deep Learning: Classification, Clustering, Regression, NLP, Computer Vision, GANs, PCA, Time Series Forecasting, ZenML, MLflow, Kubeflow, Hypothesis Testing, CNN, RNN, LSTM
Generative AI tech Stack: AI Agents, Transformers, GPT, Gemini, Llama, RAG, PEFT (LORA & QLORA) – Fine Tuning LLM, RAGAS, Langchain, Hugging Face, n8n, Moveworks Creator Studio, Google Dialogflow CX, Vertex AI
Data Visualization: Tableau, Power BI, Google Looker
Cloud Services: AWS (Bedrock, EC2, Lambda, Fargate, S3, RDS, DynamoDB, SageMaker, Redshift, Glue, Kinesis, ECS, EKS, SQS, CloudWatch, CloudFormation, Kendra), Azure (OpenAI), Databricks, Snowflake
Tools & Technologies: Excel, CI/CD-GitHub Actions, Jira, Confluence, Docker, Kubernetes

EDUCATION

The University of Texas at Dallas
Master of Science, Business Analytics (MSBA),
Specialization in Applied Machine Learning

CERTIFICATIONS & HONORS

- Scholar with Recognition – UT Dallas
- Microsoft Certified: Power BI Data Analyst
- AWS Generative AI Developer - Professional
- AWS Certified Cloud Practitioner