KAVISH DAFTARI
Boston, MA | 804-833-8108 | kavishdaftary@gmail.com | LinkedIn | GitHub

Professional Summary

AI Engineer experience in developing and deploying scalable AI applications, specializing in Generative AI, conversational Chabot's, and agentic frameworks such as LangChain, LangGraph, and Transformers. Proficient in building and validating Proof of Concepts (POCs) for systems like AI-driven customer support, recommendation engines, and anomaly detection pipelines. Skilled in fine- tuning large language models (e.g., GPT-4, T5) and implementing advanced ML algorithms across regression, classification, and time series. Experienced in integrating AI solutions into production using tools like Docker, Kubernetes, and AWS Sagemaker, with a strong focus on MLOps for robust CI/CD pipelines and model monitoring. Expertise in designing scalable, multilingual conversational AI systems with API integrations and contextual understanding. Committed to optimizing AI systems for real-world impact in industries like healthcare, finance, and e-commerce.

KEY HIGHLIGHTS AND ACHIEVEMENTS

- Designed and deployed enterprise-scale Generative AI solutions using AWS SageMaker, Bedrock, Lambda, API Gateway, and Azure OpenAI for multi-cloud compatibility.
- Spearheaded developing, deploying, and optimizing machine learning models and data-driven solutions across various cloud platforms, including AWS, Azure, and GCP.
- Led a team in creating an Enterprise Data Lake, consolidating multiple data sources, and ensuring seamless integration for advanced analytics. I directed the development of scalable AI and ML services, including predictive analytics, anomaly detection, and natural language processing (NLP) systems, using tools like Python, Docker, Kubernetes, and Kubeflow.
- Built interactive dashboards and financial data visualizations using Jupyter Notebook and Tableau, enabling stakeholders to gain real-time insights.
- Integrated Gen AI models into clinical decision-support systems.
- Experience with Agentic AI frameworks such as Lang Chain, Lang Graph, and Crew AI to develop autonomous AI systems.
- As a Data Scientist and Engineer, I developed, deployed, and optimized machine learning models and data-driven solutions across various cloud platforms, including AWS, Azure, and GCP.
- Worked on Microsoft Power Automation tool to automate Business Processes.
- Collaborated with traders, risk analysts, and finance teams to deploy predictive analytics solutions for banking and capital markets.
- Deploying large-scale AI systems, specializing in Retrieval-Augmented Generation (RAG), LLM fine-tuning, vector databases, and generative AI applications.
- Designed and developed autonomous AI agents capable of multi-step reasoning, task planning, and decision- making using LLMs, LangGraph, and reinforcement learning techniques.
- Developed and deployed machine learning models using Vertex AI, streamlining the process of training, tuning,and deploying models on Google Cloud.
- Skilled in leveraging Snowflake for large-scale data warehousing and analytics, coupled with practical deployment of ML models using the Cortex platform.
- Responsible for designing and managing end-to-end data pipelines, improving ETL processes, and optimizing data storage and processing frameworks to reduce processing times and enhance data accuracy.
- Leveraging advanced technologies and developed automated testing frameworks, CI/CD pipelines, and robust monitoring systems to ensure production-ready ML models.
- Develop AI/ML-driven predictive monitoring models using Splunk MLTK or Datadog AIOps to forecast failures and performance degradation.
- Hands-on experience integrating Kore.AI bots with enterprise systems, REST APIs, and deploying solutions on cloud environments like AWS or Azure.Experience in enhancing bot capabilities using NLP, analytics, and contextual data for improved customer engagement.
- Building and deploying AI-powered applications using Google ADK.

TECHNICAL SKILLS

Programming: SQL, Python, R, SAS, VBA, Unix Shell Scripting, Scala

ETL Tools- MS SQL Server, Azure Data Factory, AWS Glue, Apache Nifi, Talend, Informatica, SSIS

Big Data- Hadoop, Apache Spark, Hive, Cassandra, MongoDB, HBase, ElasticSearch, Redis

Python Libraries- Spacy, NLTK, Scikit Learn, Seaborn, Pandas, Numpy, Keras

NLP: Spacy, NLTK, Hugging Face Transformers, GenSim, Flair

MLOps: MLflow, DVC, Airflow, Kubeflow, Seldon Core, Git, GitHub

Professional Experience

PDI Nice Pack Inc. USA – Senior AI Engineer-With Gen AI Jan *2023– Present*

Project Outline: Developed of an enterprise Generative AI automation platform to modernize document processing and quality operations at PDI Nice Pack Inc. The solution integrated RAG-based knowledge retrieval, OCR-enhanced intelligent document processing, and LLM-driven reasoning to extract, summarize, and generate business-critical information

Responsibilities:

- Developed and deployed AI-driven applications using Google ADK, leveraging pretrained LLMs to deliver domain-specific solutions for enterprise clients.
- Designed Gen AI models to continuously analyze vast amounts of transactional data, user behavior patterns, and market trends, enabling the system to proactively identify and flag suspicious activities or outliers that might indicate fraud or other risks in real time.
- Built scalable API services for serving RAG-based generative AI solutions on AWS Sagemaker.
- Deployed models on AWS SageMaker for scalable inference and real-time predictions.
- Developed and optimized AI applications using Python, LangChain, and CrewAI, implementing Retrieval-Augmented Generation (RAG) and Agentic AI workflows for intelligent automation.
- Engineered a solution to dynamically generate SEO-friendly headlines by analyzing HTML web pages. Integrated SEMrush API and applied techniques like keyword matching, Rogue Score computation, LLM reranking, and RAG to ensure semantic relevance. Deployed the system using GCP services, enhancing search engine visibility and driving user engagement with measurable improvements in CTR and organic traffic.
- Built and optimized backend microservices using Java and Spring Boot, connecting APIs to deliver seamless data exchange between Kore.AI bots and core applications.
- Integrated vector search & embeddings using Amazon Kendra, Bedrock Knowledge Bases, and OpenSearch to improve AI-driven recommendations.
- Researched and implemented Large Language Models (LLMs) like GPT and explored Azure Microsoft Copilot Studio for automating insurance document generation, customer communication, and policy recommendations.
- Established CI/CD pipelines for AI models using AWS CodePipeline, GitHub Actions, and MLflow, enabling continuous training and deployment.
- Ensured AI security & compliance (GDPR, HIPAA, SOC 2) while mitigating risks like prompt injection, adversarial inputs, and API abuse.
- Expertise in developing autonomous agents and agentic workflows using frameworks like LangChain or AutoGen, with strong Python skills.

- Integrated Snowflake with Cortex by creating data-mart views, materialized views and staging layers, ensuring high-performance access for ML training and inference.
- Azure AI Search Indexing: Implementing advanced search indexing solutions using Azure AI to enhance data accessibility and retrieval.
- Deployed and managed Kore.AI solutions on AWS, implementing secure API integrations, real-time analytics, and enhancing overall chatbot performance and user experience.
- Developed and deployed sophisticated Generative AI applications using LangChain and LangGraph, orchestrating multi-step conversational agents and complex workflow automation, improving interactive response quality by 30%.
- Design, develop, and optimize ETL processes using SQL, ensuring data quality, accuracy, and performance.
- LLM RAG Chatbot: Support development of chatbot using RAG to improve customer support and interaction.
- Implemented advanced Gen AI prompt engineering techniques to enhance confidence scoring
- Led NLP projects in banking, utilizing text generation techniques such as GPTs and LSTM networks to improve customer service chatbots.
- Standardized cross-cloud AI security policies, ensuring real-time threat mitigation via AWS Security Hub and Azure Defender.

Environment: Python, LangChain, LangGraph, CrewAI, GEN AI, Agentic AI, OpenAI, Azure OpenAI, Azure ML, AKS, Azure Functions, FAISS, Pinecone, Weaviate, OAuth2, JWT, API Management, DevOps, Microsoft Copilot, GitHub Actions, ETL, GCP, CI/CD, Agile/Scrum.

Atmecs technologies Pvt ltd – AI Engineer August *2021 –July 2022*

Project Outline: Collaborated with engineering teams to integrate AI-driven solutions into process improvement initiatives across various business verticals, ensuring seamless alignment with organizational goals. Worked closely with cross-functional stakeholders to identify automation opportunities, translate operational challenges into scalable AI use cases, and drive end-to-end implementation.

Responsibilities:

- Experience in building, editing, testing, and deploying large scale machine learning models using AWS Sage Maker, lambda, API Gateway, and CloudWatch.
- Part of the team that performed MLOps - built and tested existing models for Performance, provide necessary support for model deployments, tune hyperparameters, bias and fairness testing, keep any eye out for outliers, scalability by monitoring various metrics on AWS CloudWatch, also evaluate the constraints for model re-training by gathering feedback on model performance and communicating the findings to Data scientists
- Other responsibilities included communicating with Stakeholders, being a scrum master, solving customer tickets, backend or frontend software development, Unit & Integration testing, code deployment and bug fixes.
- Designed & Implemented on the backend of a tool that creates & stores feedback, based on the employees Productivity & Quantity, used in warehouses across world.
- Integrated Google ADK-based AI models with business applications for intelligent document processing and automated reporting.
- Established CI/CD pipelines for ML models (versioning, automated testing, monitoring) and integrated with Snowflake for data retrieval and result storage.
- Deployed LangChain-based frameworks to orchestrate RAG flows including prompt templates, retrievers, and chains.
- Automated and scheduled computations that figure Feedback eligibility among warehouse employees which saved 3 hours of manual work/month across all warehouses globally.
- Collaborated with cloud teams to deploy Google ADK solutions on Vertex AI, ensuring scalability, monitoring, and security compliance.
- Implemented MLOps pipelines using Vertex AI, Kubeflow, and CI/CD tools, ensuring robust model deployment, monitoring, and retraining processes.
- Led the migration of 1PB+ data to Oracle Cloud Infrastructure, enhancing analytics capabilities and reducing costs.

Environment Tools – Python (Beautiful Soup, NLTK, SpaCy, Scikit-learn, Flask, PyTorch, Pandas, Numpy, Matplotlib, seaborn, plotly), SQL, Java, AI/ML, Apache Kafka, NAWS tools.

Education

Northeastern University, Boston, MA
Master of Science in Software Engineering Systems |
Relevant Coursework: Neural Modeling, Object-Oriented Design, Parallel Machine Learning, Data Science Engineering, Advances in Data Science Architecture

L.J. Institute of Engineering & Technology, Gujarat, India
Bachelor of Engineering in Information Technology |
Relevant Coursework: Artificial Intelligence, Computer Vision, Algorithms, Data Analytics, Machine Learning, Database Management Systems