Adithya Karumuri

Fort Collins, CO | +1 (970) 689 0873 | adithyakarumuri1@gmail.com |

SUMMARY

Experienced and results-driven AI/ML Engineer with 5+ years of experience designing, deploying, and scaling end-to-end machine learning and AI systems across healthcare, finance, gaming, and public sector domains. Proven track record in developing LLM-powered applications, real-time personalization systems, and predictive models for risk detection, customer behavior, and operational optimization. Adept in Python, PySpark, SQL, and cloud platforms (AWS, Azure, GCP), with hands-on expertise in MLOps, containerization (Docker, Kubernetes), CI/CD, and model monitoring using MLflow. Strong foundation in NLP, generative AI (GPT, LLaMA), time-series forecasting, and A/B testing. Known for cross-functional collaboration, stakeholder alignment, and delivering scalable, compliant, and production-ready AI solutions aligned with business KPIs.

SKILLS

Methodologies: SDLC, Agile, Waterfall
Languages: Python, R, SQL, MATLAB, HTML, CSS, Shell Scripting
Frameworks & Libraries: TensorFlow, PyTorch, Keras, Scikit-Learn, NumPy, Pandas, Matplotlib, Flask, Transformers Architecture
Big Data & ETL Tools: PySpark, Apache Spark, AWS Glue, Hadoop, MapReduce, HDFS, Zookeeper, Hive, Pig, Kafka, HBase, Spark Streaming
Databases: MySQL, PostgreSQL, SQL Server 2008, MongoDB, Cassandra, Teradata, Amazon Redshift, RDS
Data Visualization: Tableau, Power BI, AWS QuickSight, Excel (Pivot Tables, VLOOKUP)
Machine Learning & Deep Learning: Linear/Logistic Regression, Clustering, SVM, PCA, Random Forest, Boosting, Lasso, Ridge, CNN, RNN, Fine- tuning & Transfer Learning
Generative AI & LLMs: Foundation Models (GPT, LLaMA, Claude), Prompt Engineering, Tokenization (BPE, WordPiece), Reinforcement Learning from Human Feedback (RLHF)
RAG & Vector DBs: Pinecone, FAISS, Weaviate, ChromaDB, hybrid retrieval, semantic search
Compliance & FinOps: Guardrails, AI Governance, Secure AI Deployment, Cloud Cost Optimization
Cloud & DevOps: CI/CD with AWS CodePipeline, CodeBuild, CodeDeploy, CloudWatch Monitoring, Jenkins, Git, GitHub, GitLab, Cloud Architecture, Shell/Bash Scripting
Agentic AI & GenAI: CrewAI, LangChain, LangGraph, OpenAI, GPT-4, Claude, Bedrock
Tools & APIs: Postman, REST APIs, JSON Parsing, Parquet
IDEs: Jupyter Notebook, PyCharm, Visual Studio Code
Operating Systems: Windows, Linux, MacOS
Soft Skills: Problem-Solving, Communication, Collaboration

EXPERIENCE

Morgan Stanley, USA | AI/ML Engineer **Nov 2024 - Present**

- Architected LLM-powered agentic systems using LangChain, LangGraph, and CrewAI to automate workflows in finance and IT domains, implemented memory management and inter-agent communication, improving task completion rate by 30%.
- Architected and fine-tuned domain-specific Large Language Models (LLMs) using HuggingFace Transformers and LangChain, improving clinical and legal document summarization accuracy by 27%.
- Designed and deployed RAG pipelines with vector databases (Pinecone, FAISS) and OpenAI GPT-4, achieving 35% improvement in retrieval accuracy and reducing manual data lookup by 40%..
- Designed and deployed AI/ML solutions using AWS SageMaker, Bedrock, and Langchain Agents for compliance risk modeling (KYC/AML), intelligent appointment scheduling, and semantic search/summarization of regulatory filings and client communications.
- Integrated financial data into RAG (Retrieval-Augmented Generation) models and deployed GenAI-powered chatbots with Amazon Bedrock to boost digital self-service by 20% and improve customer experience across financial services.
- Collaborated with ServiceNow engineers to embed GPT-based AI into service catalogs, streamlining employee self-service and reducing first-response time by 50%.

Intex Technologies, India | AI/ML Engineer **Dec 2020 - Dec 2023**

- Led Agile-based AI/ML initiatives for enterprise IT clients, delivering end-to-end ML solutions using supervised, unsupervised, and reinforcement learning algorithms; developed XGBoost-based failure prediction models, reducing system downtime by 15%, and NLP pipelines using spaCy and BERT for enhanced root-cause analysis and automated feedback classification, improving efficiency by 20%.
- Deployed GenAI solutions on AWS Bedrock using Claude and GPT-4 for document summarization, client communication analysis, and regulatory filings, cutting processing time by 45%.
- Designed, trained, and deployed deep learning models (CNNs, RNNs, LSTMs, Transformers) for real-time anomaly detection and infrastructure monitoring; collaborated with DevOps teams to embed AI components within AIOps and observability platforms, achieving over 90% classification accuracy and reducing mean time to resolution (MTTR).
- Architected RAG-based AI assistants by integrating OpenAI's APIs with vector databases (Pinecone, FAISS) and document stores, enabling dynamic grounding of LLM responses on enterprise knowledge bases.
- Developed scalable ML pipelines with PySpark, AWS Glue, and Redshift for real-time analytics, boosting business insight delivery by 60%.
- Developed LLM-powered knowledge assistants within ServiceNow using its Virtual Agent and Integration Hub to automate ticket triage, incident categorization, and knowledge article generation.

Groovy Web, India | ML Engineer **Jan 2019 - Nov 2020**

- Built and deployed LLM-driven assistants for ITSM platforms using OpenAI APIs and ServiceNow's Virtual Agent; automated ticket triage and knowledge base generation.
- Established MLOps pipelines for real-time model deployment and monitoring , integrated AI observability tools for early anomaly detection
- Transformed and cleansed large-scale datasets using SQL, PL/SQL, and SAS, optimizing ETL workflows for business intelligence reporting; built automated data pipelines in Python (Pandas, NumPy) to support scalable model training, feature engineering, and data validation.
- Developed and deployed supervised machine learning models (Random Forest, SVM, Naïve Bayes, KNN), utilizing Grid Search and feature selection for optimization, and achieved high F1-scores and ROC-AUC in classification tasks across multiple domains.
- Created advanced data visualizations and interactive dashboards using Matplotlib, Seaborn, Plotly, and Power BI, and conducted detailed exploratory data analysis (EDA) to reveal patterns, trends, and anomalies driving key strategic decisions.
- Built time-series forecasting models using DeepAR and XGBoost for no-show prediction and supply chain optimization, improving planning accuracy by 22%.

EDUCATION

Master's in Computer Information Systems | Colorado State University, Fort Collins, CO **May 2025**
Bachelor of Technology in Mechanical Engineering | M.V.G.R College of Engineering, Vizianagaram, India **Jun 2020**

CERTIFICATIONS

Mastering Generative AI with OpenAI, LangChain, and LlamaIndex V2 (Ineuron.ai)
Microsoft Azure Machine Learning (Microsoft)
AWS Certified Solution Architect (AWS)