Bhanu Prakash Reddy, Marreddy

bhanuprakash2511@gmail.com (214) 218-6957 Dallas, Texas LinkedIn

SUMMARY: Data Scientist & AI Engineer with 8 years of expertise in Generative AI, Agentic AI, and MLOps. Proficient in LangGraph, Google ADK, LangChain, RAG architectures, and LLM fine-tuning (Gemini, GPT-4). Strong foundation in Python, SQL, PySpark, FastAPI, ensemble ML (XGBoost, Random Forest, SVM), deep learning (LSTMs, Transformers), and advanced NLP (BERT, semantic search, vector embeddings). Experienced in cloud-native deployments (OpenShift, Kubernetes), CI/CD pipelines, asyncio concurrency, and production monitoring (Prometheus, Grafana). Skilled in statistical analysis, AB testing, time series forecasting, and business intelligence tools.

PROFESSIONAL WORK EXPERIENCE:

Citigroup Inc., Dallas, USA

Sr. Generative AI Engineer (Python, SQL, Google ADK, LangGraph, FAST API OpenShift/Kubernetes) Apr 2025 – Present

- AI/ML & Platform Engineering: Engineered and deployed a Hybrid AI Engine (Traditional ML and Gemini 2.5 Pro GenAI) to automate critical regulatory report tagging and rationale generation. Scaled the solution to support 26 high-value financial reports (18 Risk reports via ML, 8 Finance reports via GenAI), processing complex instruction sets up to 750 pages (10 MB).
- Optimized Traditional ML performance for 18 risk reports, conducting rigorous experimentation with Random Forest, XGBoost, SVM, and Logistic Regression. Deployed tailored models with performance-enhancing techniques like SMOTE, achieving a recall of about 0.78.
- Advanced GenAI solution design for 8 Finance reports, implementing advanced patterns including Retrieval-Augmented Generation (RAG) and similarity-based few-shot learning. Final deployment on Gemini 2.5 Pro resulted in a 70% improvement in automated rationale quality and interpretability.
- Drove end-to-end MLOps and Deployment on Openshift, configuring deployment.yaml, cronjob.yaml, and values.yaml for automated scheduling and scaling. Architected and exposed the model via multiple high-performance FastAPI endpoints.
- Led LLM-based COBOL code generation POC using Gemini 2.5 Pro, transforming 5,000+ data mapping rules into executable COBOL within 1 week (vs. 10-week manual estimate), demonstrating 50x velocity improvement. Designed multi-stage LLM pipeline with complexity-based routing (80% direct translation, 20% chain-of-thought decomposition), achieving 92% first-pass compilation and 98% functional equivalence on 500K+ test records.
- Deployed production Agentic AI systems using Google ADK and LangGraph to automate daily trade settlement validation (2,000+ trades/day) and AML product approval reviews, achieving 95% automation rate. Built ReAct agent framework with specialized agents for data extraction, rule validation, and anomaly detection; integrated 45+ compliance checks reducing settlement validation from 4 hours→15 minutes daily.

Globe Life Insurance, Dallas, USA

Sr. Data Scientist (Gen AI, Claude, Python, SQL, Tableau, QuickSight, AWS Glue, Redshift) Apr 2024 – Mar 2025

- Developed a RAG chatbot for Globe Life Insurance – Sales Team using fine-tuned Claude models on AWS Cloud enabling the content creation, Question Answering, text to text generation, Classification and optimizing language understanding and generation for enhanced user interactions.
- Applied chunking strategies to effectively process and analyze large blocks of text for more coherent and contextually relevant responses.
- Utilized vector representations of text and semantic search techniques to enhance the accuracy and speed of query processing, leading to more satisfying user experiences.

Metamorphix Inc, Remote, USA

Data Scientist (NLP, Python, SQL, Pyspark, ADF, Databricks) Jul 2023 – Dec 2023

- Developed ensemble models using NLP techniques, achieving a 20% performance improvement through hybrid feature engineering module, sentiment aware embedding layer (BERT) and error analysis module.
- Designed scalable data pipelines using Apache Kafka and Spark Streaming processing 500GB+ daily for low latency real-time sentiment analysis at scale.

- Validated context-aware recommendation engines that drove a 20% increase in user engagement on Cascades Platform by incorporating real-time user data such as location, time of day, and device type to deliver personalized recommendations.
- Collaborated across departments to implement ML-driven product features benefiting 70,000+ active users at Metamorphix - Cascades Platform.

Yoamigos Webservices (Clients: TESCO, PEPSICO, Claris Health), India **Jun 2017 – Jul 2022**
Senior Data Scientist (NLP, Timeseries, Classification, Regression)

- Detecting Industry Trends on social media using NLP: Processed and analyzed extensive data, including 2+ million unique F&B products in 125+ categories. Utilized techniques like parsing, lemmatizing, and POS tagging, to process tweets. Identified precise topics and detected trends accurately using Bayesian classifiers, fuzzy matching algorithms, and Latent Dirichlet Allocation models.
- Forecasted sales for more than 300 products across 5 retailers in multiple countries resulting in trade promotion optimization (TPO). The built Auto Regressive Distributed Lag (ARDL) model achieved 13% WMAPE that far exceeded the accuracy of manual forecasts.
- Deployed and productionized LightGBM and Ordinary Least Squares models on Azure platform, trained on data normalized data using ALS factorization, to recommend product placement for retail stores in various European markets on a weekly basis.

Data Scientist

- Designed a payment integrity solution to classify claims into different levels of risk using Support Vector Classifier (SVC) with a recall of 91% at AUC of 0.95. Integrated the SVC model into the existing software development framework. Reduction in false positives increased the efficiency of auditors by 40%.
- Customer segmentation using RFMT model: Applied the Recency, Frequency, Monetary, and Time (RFMT) model to analyze patient data. Employed K-Means, Gaussian, and DBSCAN algorithms to classify patients into distinct groups. Conducted cluster factor analysis using methods such as elbow, dendrogram, silhouette, Calinsky–Harabasz, Davies–Bouldin, and Dunn index to ensure the meaningfulness of patient groupings. Implemented the majority voting (mode version) technique to select the most relevant patient clusters.
- Increased customer lead utilization by 20% by using Random Forest to classify untouched leads as potential customers or not. Achieved accuracy of 98% and F2 score of 35%.

ACADEMIC PROJECT EXPERIENCE

- Architected hybrid agentic system combining research and coding agents using Google ADK + LangGraph, orchestrating 6 specialized agents (planner, researcher, coder, tester, critic, synthesizer) for autonomous technical report generation with executable code examples. Implemented hierarchical planning with tool use: research agents performed parallel literature search and extraction (arXiv, Semantic Scholar APIs), coding agents generated validated implementations, critic agents provided iterative feedback via self-reflection loops using Google ADK's tool-calling framework. Engineered shared memory architecture with vector stores (FAISS) and LangGraph state machines for cross-agent context sharing and consensus-based decision making; reduced hallucinations by 40% through multi-agent verification and structured output validation. Achieved 78% end-to-end task completion on complex queries requiring research synthesis + code generation, with ROUGE-L 0.81 for summaries and 85% functional correctness on generated code vs. 55% single-agent baseline

CERTIFICATIONS

Google Cloud Certified Professional Machine Learning Engineer (Series ID: 4158) May 2023 – May 2025

VOLUNTEER EXPERIENCE

National Service Scheme, President Aug 2015 – Jul 2016

EDUCATION

- **Master of Science, Business Analytics (Data Science Track) Dec 2023**
- **Bachelor of Engineering, BITS Pilani, India Aug 2017**