JAYAPRAKASH YADAV GUNTUMANI

Houston, TX | +1 (713)-551-6420 | jayaprakashguntumani@gmail.com | LinkedIn | GitHub

Summary

AI/ML Engineer with 2+ years of experience building production ML and LLM-based systems across cloud platforms, specializing in time-series forecasting, RAG pipelines, fine-tuning, and agentic AI workflows using Python.

Projects

Role Based Access Control (RBAC) Chatbot | Link

- Designed and deployed an RBAC-enforced RAG chatbot on Google Cloud Run, implementing role-based document filtering to securely serve department-specific knowledge across 5 business units using FastAPI, LangChain, Gemini, and ChromaDB.
- Improved multi-hop reasoning quality by 42% (2.57 → 3.65) while reducing latency by 25% (5.43s → 3.56s), by optimizing chunk size, top-k retrieval, model selection, and hybrid retrieval strategies across RAG experiments.
- Built an automated MLflow evaluation pipeline to execute and compare 58 RAG configurations end-to-end, orchestrating document ingestion, inference, LLM-based scoring, and metric aggregation without human labeling.

Safe Space Agentic AI Assistant | Link

- Built AI mental health chatbot with LangGraph ReAct agents integrating 3 specialized tools (MedGemma LLM, Twilio emergency calls, Google Maps therapist finder) for 24/7 emotional support.
- Developed multi-channel AI agent system using LangChain and FastAPI with tool-augmented capabilities for crisis detection and location-aware therapist recommendations, demonstrating production LLM deployment with real-time API integrations.

Code-to-Comment Generator | Link

- Built an automated code-to-docstring generation system fine-tuned on 21K CodeSearchNet samples, achieving 0.73 BERTScore F1 (8% improvement over base CodeT5) using LoRA for parameter-efficient training.
- Reduced trainable parameters by 95% (2M vs 220M) while preserving generation quality, by applying Low-Rank Adaptation (LoRA).

Professional Experience

Tesla Solutions **Sep 2025 - Present**
Machine Learning Intern *Houston, TX*

- Designed and trained a multivariate neural network–based time-series classifier in PyTorch for well-state detection across 10+ oil & gas wells, achieving 94% accuracy and reducing transition-period misclassification by 15% compared to baseline models.
- Improved overall model accuracy by 10% by engineering choke-based temporal features (differentials, rolling standard deviation) and augmenting underrepresented transition-state samples to address class imbalance.
- Productionized the ML pipeline to reduce batch inference time by over 50% on 100K+ records by migrating preprocessing logic to C# and serving the ANN via ONNX, eliminating API and cross-language overhead.

Optum Global Solutions (UHG) **May 2022 - Jul 2024**
Associate Data Scientist *India*

- Architected a PySpark-based model monitoring pipeline to detect data, performance, and concept drift at scale, achieving a 40% processing speedup over legacy Pandas workflows.
- Streamlined the model development lifecycle by automating lag selection and seasonality detection using ACF/PACF analysis, reducing manual feature engineering effort by 25%.
- Improved forecasting accuracy by 20% by transforming time-series data into a supervised learning problem, combining Prophet-extracted trend and seasonality features with lagged variables and training an XGBoost regression model.
- Reduced report turnaround time by 5× during a Power BI migration by generating JSON configuration files via VBA, simplifying a complex 20+ step Power Automate workflow into a single scalable action.

Technical Skills

- **AI & LLMs**: RAG, Fine-Tuning, Prompt Engineering, Agentic Workflows, LangChain, Vector Databases (FAISS, ChromaDB)
- **ML & Deep Learning**: PyTorch, Time-Series Modeling, Computer Vision, NLP, Supervised/Unsupervised Learning
- **Programming & Tools**: Python, SQL, C#, FastAPI, PySpark, Numpy, Pandas, HuggingFace Transformers, Git
- **Cloud & MLOps**: GCP (Vertex AI, Cloud Run), AWS (SageMaker, Lambda), Docker, Kubernetes, MLflow, ONNX Runtime

Education

University of Houston **Aug 2024 - May 2026**
Master of Science, Engineering Data Science *Houston, Texas*

- **GPA:** 3.92/4.0
- **Coursework:** Natural Language Processing, Deep Learning for Engineers, Machine Learning, Introduction to Cloud Computing

Certifications

- Oracle Cloud Infrastructure 2025 Certified Generative AI Professional
- Oracle AI Vector Search Certified Professional