Yashwanth Reddy

AI Engineer

yashwanthreddy3047@gmail.com| +1 (848) 328-0485 | Jersey City, NJ | linkedin.com/in/Yashwanth-reddy-boddireddy

SUMMARY

AI Engineer with 4+ years of experience building production LLM applications, RAG systems, and scalable ML inference services. Deployed high-throughput AI pipelines serving ~60K requests/day on Kubernetes with GPU infrastructure. Hands-on with Python, PyTorch, LangChain, vector databases, and end-to-end model deployment on AWS.

PROFESSIONAL EXPERIENCE

AI Engineer, JPMorgan Chase & Co., New Jersey, USA 07/2025 - Present

- Built scalable inference services handling ~60K requests/day on Kubernetes with GPU-backed infrastructure, serving real-time predictions across multiple financial business units.
- Reduced model latency from ~900ms to ~580ms (35%) through request batching strategies and TensorRT optimization on production serving infrastructure.
- Developed RAG pipelines combining BM25 + FAISS with Reciprocal Rank Fusion and Cohere reranking for anomaly detection and enterprise data insights, improving answer relevance from 82% to 94%.
- Deployed and maintained production ML services through CI/CD pipelines with automated quality evaluation using Ragas, collaborating with cross-functional teams to ensure uptime.

AI Engineer, Accenture, Hyderabad, India 10/2020 - 08/2023

- Developed NLP classification and recommendation systems for enterprise applications, improving user engagement by 18% through personalized content ranking models.
- Designed end-to-end ML pipelines for model training, validation, and deployment, reducing model release cycles from weeks to days across 6+ production solutions.
- Built document search and Q&A systems using transformer-based architectures, enabling semantic retrieval across 500K+ enterprise documents.
- Mentored 4 junior engineers on ML development practices and contributed to delivery of production-grade AI solutions across client engagements.

PROJECTS

Production RAG System for Financial Document Q&A | *Python, LangChain, FAISS, Cohere Rerank, FastAPI, Ragas, GitHub Actions, AWS*

- Built a production RAG pipeline using BM25 + FAISS with Reciprocal Rank Fusion and Cohere reranking, improving answer relevance from 82% to 94% with enforced citation grounding.
- Deployed a FastAPI streaming service on AWS (Docker, S3-backed FAISS) with sub-2s latency and a CI pipeline using GitHub Actions with Ragas-based quality evaluation.

LLM Monitoring and Observability Dashboard | *Langfuse, Prometheus, Grafana, Python, FastAPI, GitHub Actions*

- Built an LLM observability system using Langfuse, Prometheus, and Grafana to track latency, token cost, and RAG quality metrics, reducing reranking latency from 340ms to 180ms.
- Cut token cost by 22% through optimized context selection and enforced quality and latency thresholds in CI/CD using GitHub Actions.

EDUCATION

Master of Science in Data Science, New Jersey Institute of Technology (NJIT), Newark, USA 09/2023 - 05/2025

TECHNICAL SKILLS

Languages & Frameworks: Python, PyTorch, TensorFlow, Hugging Face, LangChain, FastAPI

LLM & NLP: Large Language Models (LLMs), RAG, Prompt Engineering, Transformers, NLP, Cohere Rerank

MLOps & Infrastructure: Docker, Kubernetes, MLflow, Airflow, CI/CD, GitHub Actions, Model Deployment, TensorRT

Cloud Platforms: AWS (EC2, S3, Lambda, SageMaker), Azure, GCP

Data & Storage: Apache Spark, Kafka, ETL Pipelines, SQL, Vector DBs (FAISS, Pinecone)

Monitoring & Eval: Langfuse, Prometheus, Grafana, Ragas