

VIVEK KRISHNA REPALA

+1 470 636 9622 — vivekkrishnajob@gmail.com — LinkedIn — GitHub

OBJECTIVE

Machine Learning Engineer with hands-on experience building and deploying **LLM-powered applications** and intelligent ML systems. Skilled in **model integration, prompt engineering, retrieval-augmented generation (RAG), and agentic workflows**, with a strong focus on delivering scalable, production-ready AI solutions. Seeking to apply machine learning expertise to build impactful, AI-driven features that enhance real-world operational and user experiences.

EDUCATION

Master of Science in Data Science

University of Alabama at Birmingham

GPA: 3.6

TECHNICAL SKILLS

- Programming:** Python, JavaScript, TypeScript, REST APIs.
- LLMs & GenAI:** Prompt engineering, RAG, agentic workflows, tool calling.
- LLM Frameworks:** LangChain, LangGraph, LlamaIndex.
- Backend Development:** FastAPI, Node.js, API integration.
- Frontend Exposure:** React fundamentals, UI-API integration.
- Cloud & Deployment:** AWS, Docker, CI/CD pipelines.
- Data & Retrieval:** Vector databases, embeddings, external knowledge integration.
- AI Quality & Safety:** Guardrails, evaluation pipelines, latency and cost optimization.

EXPERIENCE

Machine Learning Engineer – ReplyQuickAI (Remote)

Oct 2025 – Present

- Designed and deployed **LLM-powered applications** supporting AI agents and automated decision workflows.
- Built **RAG pipelines** integrating external knowledge sources using LangChain and vector databases.
- Implemented **agentic workflows** with tool calling and multi-step reasoning using LangGraph.
- Developed backend APIs (FastAPI) to integrate AI features with web and internal applications.
- Optimized LLM inference for latency, cost, and reliability in production environments.
- Collaborated closely with product managers and engineers to ship customer-facing AI features.

Machine Learning Intern – Honeywell

Jan 2025 – Jun 2025

- Built ML-backed services supporting analytics and operational intelligence platforms.
- Assisted in deploying AI models into production systems with monitoring and reliability checks.
- Worked with structured and time-series data to support predictive and diagnostic use cases.

Data Scientist – Amadeus

Dec 2021 – Dec 2023

- Developed scalable analytics and ML solutions processing **50M+ records**.
- Built backend data services and APIs supporting analytics-driven applications.
- Partnered with engineering teams to productionize models and data pipelines.

PROJECTS

AI Agent Platform with LLM Orchestration

- Built a production-ready AI agent system using **LangChain and LangGraph**.
- Implemented retrieval-augmented generation with external data sources.
- Integrated AI agents into backend services via REST APIs.

LLM-Powered Knowledge Assistant

- Developed an LLM-powered assistant to answer domain-specific queries.
- Designed prompt templates, evaluation metrics, and guardrails to reduce hallucinations.
- Optimized response latency and token usage for cost-efficient deployment.

Cloud-Deployed AI Application

- Containerized AI services using Docker and deployed on AWS.
- Implemented CI/CD pipelines for safe, repeatable deployments.
- Added monitoring for availability, latency, and failures.