

Harish Bejawada

United States

+1 (716) 335-1329 | b.harish2727us@gmail.com | [LinkedIn](#) | [GitHub](#) | [Personal AI Agent](#)

SUMMARY

Seeking AI Agent Engineer where I have 3+ years of experience designing and deploying agent-based systems using LLMs, tools, memory, and planning for enterprise automation. Proven expertise in workflow orchestration, multi-step workflows, RAG pipelines, and human-in-the-loop safety patterns, delivering production-grade solutions with measurable accuracy, reliability, and task success. Skilled in Python, LangChain, Semantic Kernel, vector databases, and evaluation frameworks, with a strong record of cross-functional collaboration and software engineering excellence from prototype through production deployment.

WORK EXPERIENCE

Moody's Analytics

United States

Software Engineer | AI-Driven Financial Intelligence Platform

Nov 2025 – Present

- Designed and deployed a multi-agent orchestration system using **LangGraph** and **AWS Bedrock (Claude)**, where planner, retrieval, and analyst agents collaborate via tool use and function calling to answer complex regulatory queries, reducing analyst research time by 60%.
- Built a production-grade **RAG** pipeline with **LlamaIndex**, **FAISS**, and hybrid (semantic + BM25) retrieval over 10k+ financial regulatory documents, with source citation and grounding checks that cut hallucination rates by 45% in evaluations.
- Implemented **human-in-the-loop** guardrails and evaluation frameworks (LLM-as-judge, golden datasets, trajectory scoring) with approval workflows for high-risk actions, enabling safe autonomous execution of compliance processes.
- Worked cross-functionally with enterprise client teams to scope use cases, run pilot deployments, and iterate on prompts, tools, and integrations, taking 3 agentic use cases from prototype to production in under 8 weeks each.
- Developed **MLOps/LLMOps** orchestration with **AWS SageMaker Pipelines** and **MLflow** automated retraining, prompt/version management, and A/B testing reducing deployment drift incidents by 70% and ensuring reliability.
- Engineered low-latency serving infrastructure (**FastAPI**, **Redis** feature store, streaming responses) cutting end-to-end agent response latency by 55% for real-time client-facing workflows.

Cognizant Pvt Ltd.

India

AI/ML Engineer | Applied NLP & Production ML Systems

Feb 2022 – Jan 2024

- Fine-tuned **BERT** and **RoBERTa** encoders for multi-class transaction intent classification across 40+ categories, lifting macro-F1 from 0.78 to 0.92 and retiring a legacy keyword-matching service.
- Built a semantic search layer over merchant and case records using **Sentence-Transformers** embeddings and approximate nearest-neighbor indexing, letting analysts surface similar historical cases in under 200ms instead of manual keyword lookup.
- Developed a document understanding pipeline combining **OCR**, layout-aware parsing, and **NER** to extract structured fields from unstructured dispute submissions, automating intake for roughly 4k documents per week.
- Trained and shipped a gradient-boosted risk scoring model (**LightGBM**) on 200+ engineered behavioral and velocity features, holding 99.8% precision at production volume while cutting analyst false-positive load by 30%.
- Designed the model serving stack **FastAPI** inference services, **Redis** feature caching, **Docker** on **AWS ECS** sustaining sub-100ms p95 latency across 10k+ daily scoring requests.
- Instrumented drift and performance monitoring (**Evidently**, **Plotly Dash**) with automated alerting on embedding drift and score distribution shifts, catching two silent degradations before they reached business metrics.
- Established the team's retraining and evaluation workflow in **MLflow** experiment tracking, model registry, champion/challenger comparison reducing the cycle from a new dataset to a validated deployed model from three weeks to four days.

Gandhi Institute of Technology and Management

India

Junior Data Scientist | Academic Analytics & Enrollment Forecasting

Mar 2021 – Feb 2022

- Built a time-series forecasting model (**Prophet**, **SARIMA**) predicting student enrollment and course demand with 87% accuracy, enabling proactive resource allocation and planning for academic operations.
- Automated **ETL** pipelines (Python, SQL) integrating records from 5+ legacy databases into a unified analytical warehouse, supporting multi-step workflows for data-driven decisions.
- Presented EDA-driven retention insights (**Seaborn**, **Matplotlib**) directly to academic leadership, shaping two policy changes on at-risk student outreach through cross-functional collaboration.
- Designed evaluation frameworks to measure model performance against baseline metrics, ensuring reliability and task success across forecasting and recommendation systems.

PROJECTS

Autonomous Multi-Agent Document Intelligence System

LangGraph, FAISS, Claude API, AWS

- Built a multi-agent system (router, retriever, analyst, and critique agents) that autonomously ingests, chunks, and indexes 5k+ PDFs and answers complex cross-document questions with cited sources.
- Implemented tool use and structured function calling so agents can query **FAISS**, run calculations, and invoke web search, with a critique agent verifying claims before final output.
- Combined hybrid search (semantic + BM25) with reranking to improve retrieval relevance by 30% over pure vector search; grounding and few-shot prompting reduced hallucination rates by 45% in evaluations.
- Deployed a **Streamlit** frontend with chat history and agent-trace visualization as a serverless app on **AWS Lambda** and API Gateway.

Enterprise Workflow Automation Agent

LangChain, FastAPI, PostgreSQL, Docker

- Designed an agent-based workflow automation system that orchestrates multi-step processes across enterprise APIs, reducing manual task completion time by 50% for operations teams.
- Integrated human-in-the-loop approval checkpoints for critical actions, ensuring safety and compliance while maintaining autonomous execution for routine tasks.
- Built evaluation frameworks measuring accuracy, reliability, and task success across 200+ test scenarios, achieving a 94% task completion rate in production.
- Deployed as containerized microservices with **Docker** and orchestrated via **Kubernetes**, supporting 10k+ daily API calls with 99.9% uptime.

RAG-Powered Knowledge Assistant

Pinecone, OpenAI API, Python, Streamlit

- Developed a **RAG** pipeline using the **Pinecone** vector database and **OpenAI** embeddings to power a knowledge assistant answering questions across 50k+ internal documents with cited sources.
- Implemented multi-step retrieval with query rewriting and reranking, improving answer accuracy by 35% over baseline retrieval approaches.
- Integrated memory and conversation history to enable context-aware follow-up questions, enhancing user experience and task success rates by 28%.
- Deployed with a **FastAPI** backend and **Streamlit** frontend, including monitoring dashboards for retrieval quality and user feedback loops.

TECHNICAL SKILLS

Agentic AI & LLMs:	LangGraph, LangChain, Semantic Kernel, LlamaIndex, Function Calling / Tool Use, MCP, Multi-Agent Orchestration, Multi-Step Workflows, Planning, Memory, Prompt Engineering, RAG, Agent Evals (LLM-as-judge), Guardrails, Human-in-the-Loop, OpenAI & Anthropic APIs, AWS Bedrock
AI & ML:	PyTorch, TensorFlow, Scikit-learn, Hugging Face Transformers, XGBoost, LightGBM, NLP, Time-Series, Anomaly Detection, Applied Generative AI
MLOps & Big Data:	AWS SageMaker, MLflow, Kubeflow, Apache Spark (PySpark), Docker, Kubernetes, GitHub Actions, A/B Testing, Model Drift Monitoring, Production Deployment, Monitoring
Vector DBs & Search:	FAISS, Pinecone, Weaviate, Elasticsearch, Redis (vector), Hybrid Search (BM25 + semantic), Vector Databases
Languages:	Python, SQL, TypeScript, Java, JavaScript, Bash
Backend & APIs:	FastAPI, Node.js, RESTful APIs, GraphQL, WebSockets, Streaming/SSE, API Integration
Cloud & DevOps:	AWS (EC2, S3, Lambda, RDS, Bedrock, ECS), GCP, Terraform, Jenkins
Databases:	PostgreSQL, MySQL, MongoDB, Redshift
Core Competencies:	Agentic System Design, Workflow Orchestration, Evaluation & Reliability, Software Engineering Excellence, Cross-Functional Collaboration, Scalable Solutions, Secure Solutions, Client Discovery & Scoping, Pilot-to-Production Delivery, Stakeholder Demos, Rapid Prototyping, System Design, Agile

EDUCATION

University at Buffalo, The State University of New York

Buffalo, NY

Master of Science in IE Data Analytics

Jan 2024 – Apr 2025