# Ravi Kumar Rayapalli

California, USA | +1 7065907391 | ravirayapalli30@gmail.com | linkedin.com/in/ravi-rayapalli | github.com/ravi-30

## Summary

- AI Engineer with 8 years of experience specializing in LLM systems, agentic architectures, and retrieval-augmented generation. Experienced in building production-grade multi-agent workflows, knowledge graphs, and hybrid retrieval systems across cloud environments. Strong background in backend engineering, data platforms, and scalable system design.

## Experience

**Optum (UHG)** Feb 2026 – Present
*AI Engineer* *USA*

- Built Legal AI systems for Optum Healthcare leveraging LLM driven document intelligence, including entity extraction, clause classification, semantic search, and automated legal workflow orchestration across large scale corpora.
- Designed and implemented agentic workflows using Google ADK, defining planner executor architectures with structured prompt chaining, dynamic routing, tool invocation layers, and coordinated multi agent execution with shared state and memory.
- Developed retrieval driven agent architectures with a dedicated retrieval agent that dynamically queries hybrid data sources, including vector stores, Neo4j graphs, and enterprise data systems, to ground responses in relevant legal context.
- Built MCP compliant tools and services exposing reusable capabilities over BigQuery, REST APIs, and internal services, enabling controlled tool invocation, schema aware querying, and seamless integration into agent execution pipelines.
- Designed and implemented legal domain knowledge graphs in Neo4j, modeling entities such as contracts, clauses, obligations, and parties, along with relationship semantics to enable dependency tracking and graph based reasoning.
- Implemented hybrid retrieval pipelines combining dense embeddings and graph traversal, where node level vector embeddings in Neo4j are used alongside relationship aware Cypher queries, orchestrated through ADK retrieval agents for context enriched reasoning.
- Enabled agent to agent communication using structured messaging protocols, for example a planning agent delegating clause extraction to a document agent and routing extracted obligations to a compliance agent, improving coordination, context propagation, and execution efficiency.
- Built evaluation and monitoring pipelines for agent systems, incorporating automated metrics such as answer correctness, retrieval precision and recall, grounding fidelity, latency, and failure analysis.
- Established DeepEval-driven CI/CD pipelines to validate LLM outputs against predefined benchmarks, including context recall, answer relevance, hallucination detection, and regression testing prior to deployment, reducing hallucinations by 25%.
- Containerized microservices using Docker, enabling reproducible environments, dependency isolation, and scalable deployment across development and production systems.
- Leveraged Google Cloud platform including Cloud Run and GKE for scalable execution, BigQuery for analytics, Cloud Storage for data persistence, Pub Sub for event driven workflows, Cloud Functions for serverless execution, and Cloud Build with Artifact Registry for CI CD and artifact management, secured via IAM.

**Talent Screen** Jul 2025 – Feb 2026
*AI Engineer Intern* *USA*

- Built Maya Talent Agent using LangGraph, implementing a stateful, graph-based conversational system for candidate intake, screening, question answering, and interview orchestration across the hiring lifecycle, reducing manual effort by 70%.
- Designed an end to end screening pipeline with LangGraph state machines, capturing candidate inputs, parsing resumes, executing structured evaluation stages, generating responses, scheduling interviews, and maintaining a persistent candidate profile.
- Implemented an intent classification layer using LLM based routing and lightweight classifiers to direct requests across resume ingestion, screening, and FAQ pipelines, enabling controlled task specific execution.
- Developed a resume ingestion pipeline using Unstructured.io and custom parsing logic to extract, normalize, and structure candidate data into canonical schemas for downstream processing.
- Built a Resume Parser Tool leveraging LLM extraction and schema validation to identify entities such as skills, work experience, education, and role specific qualifications, with post processing for normalization.
- Implemented a Skill Matcher Tool using embedding similarity and rule based scoring to align candidate profiles with job requirements, combining semantic matching with structured constraint checks.
- Engineered a multi stage screening workflow in LangGraph, including profile construction, candidate scoring, gap analysis, follow up generation, and interview scheduling, enabling adaptive and context aware interactions.
- Developed a gap detection module using LLM reasoning and rule based validation to identify missing qualifications, ambiguous experience, and incomplete profile attributes.
- Designed dynamic follow up generation using prompt templates and contextual memory to iteratively collect missing candidate information during multi turn conversations.
- Integrated scheduling capabilities via external calendar APIs, enabling automated interview coordination and workflow progression within the agent pipeline.

- Built an Embedding Search Tool using Sentence Transformers and vector databases such as FAISS, Milvus, and ChromaDB to support retrieval augmented generation for candidate context and FAQ responses.
- Implemented a RAG based question answering pipeline combining FAQ document retrieval with LLM grounded response generation for candidate queries about roles, hiring process, and policies.
- Designed state management and memory handling using LangGraph state stores and conversation buffers to maintain candidate context, track intermediate outputs, and ensure consistency across multi step workflows.
- Built MCP compliant components including a centralized tool registry for managing tool schemas, invocation policies, and execution handlers for services such as resume parsing, skill matching, scheduling, and retrieval.
- Implemented guardrails including schema validation, input sanitization, timeout control, retry strategies, and LLM fallback mechanisms to ensure robust execution under tool failures and incomplete responses.
- Improved system reliability through controlled tool orchestration, explicit state transitions, and fault tolerant execution patterns across LangGraph nodes.
- Built evaluation frameworks using RAGAS, BERTScore, and LLM as a judge, combined with human in the loop validation to measure answer relevance, grounding, and consistency in production.
- Established observability using structured logging, distributed tracing, and retrieval level analytics dashboards to monitor latency, tool performance, retrieval quality, and model drift.
- Developed end to end RAG pipelines with document ingestion via Unstructured.io, adaptive semantic chunking strategies, and embedding generation using Sentence Transformers across heterogeneous document sources.
- Engineered hybrid retrieval systems combining BM25 sparse retrieval with dense vector search across FAISS, Milvus, and ChromaDB, using reciprocal rank fusion and cross encoder reranking to optimize precision and recall for complex queries.

**Equiply.io** Jan 2025 – Mar 2025
*Software Engineer Intern* *USA*

- Developed backend services using Node.js and TypeScript to handle document ingestion, processing, and data indexing workflows.
- Built and maintained RESTful APIs to support data retrieval, search functionality, and internal application integrations.
- Designed scalable API layers and middleware with request validation, error handling, and structured logging to ensure reliability and maintainability.
- Implemented search functionality using vector similarity and indexing techniques, improving relevance and performance of data retrieval systems.
- Optimized backend performance through batching, caching strategies, and efficient request handling.
- Developed reusable modules and services to support data processing and integration across multiple internal tools.
- Collaborated with cross-functional teams to design and integrate backend services with frontend applications.

**EY India** Dec 2020 – Jul 2023
*Senior Software Engineer* *India*

- Led development of distributed backend services and data platforms supporting high-volume enterprise applications.
- Architected and implemented scalable microservice systems using containerized deployments and cloud infrastructure.
- Built large-scale data pipelines to process and analyze structured and unstructured datasets for analytics and internal tools.
- Introduced machine learning integration into backend systems, enabling predictive analytics and intelligent automation features.
- Designed service monitoring, logging, and performance tuning strategies to improve system reliability and observability.
- Mentored junior engineers and conducted technical design reviews, code reviews, and architecture discussions.
- Improved development efficiency by establishing CI/CD pipelines, automated testing frameworks, and deployment workflows.
- Collaborated with data science teams to integrate ML models into production applications through API-based services.

**Aituristic Software Pvt Ltd** Feb 2018 – Nov 2020
*Software Engineer* *India*

- Designed and developed scalable backend services and APIs using Python/Java to support enterprise web applications handling large datasets.
- Built data processing pipelines to ingest, transform, and store application data from multiple internal and external sources.
- Implemented microservice-based architecture enabling modular service deployment and improved system scalability.
- Optimized SQL queries and database indexing strategies to improve application performance and reduce query latency.
- Integrated third-party APIs and messaging systems to enable real-time data exchange between services.
- Developed automation scripts and background jobs for data validation, monitoring, and scheduled processing tasks.
- Contributed to CI/CD pipelines and containerized application deployments using Docker and build automation tools.
- Worked closely with product and frontend teams to deliver end-to-end application features and API integrations.

**Innovapath IT Solutions Pvt Ltd** May 2017 – Jan 2018
*Software Engineer* *India*

- Developed backend services in Python and Java supporting internal web applications and data processing workflows.
- Assisted in building REST APIs and microservices for internal tools using frameworks such as Flask/Spring Boot.
- Implemented SQL queries and database schemas to support application features and reporting dashboards.
- Built data ingestion and preprocessing scripts to transform and clean structured and semi-structured datasets.
- Participated in unit testing, debugging, and code reviews, improving code reliability and maintainability.
- Collaborated with senior engineers using Git-based version control and Agile development practices.

## Education

**Cleveland State University** USA
*Master of Science in Information Systems* *May 2025*

**Jawaharlal Nehru Technological University** India
*Bachelor of Technology in Electronics and Communication Engineering* *Jan 2018*

## Technical Skills

- Programming Languages: Python, SQL, JavaScript, TypeScript, Java
- AI & LLM Frameworks: Google ADK, LangChain, LangGraph, LlamaIndex, Hugging Face Transformers, Pydantic
- Machine Learning: PyTorch, Scikit-learn, MLflow, Feature Engineering, Model Evaluation
- RAG & Retrieval Systems: Vector Databases, Milvus, Semantic Search, Embeddings, Hybrid Retrieval, BM25
- LLM Platforms: Vertex AI (Gemini), OpenAI, Anthropic Claude, LLAMA, Mistral, AWS Bedrock
- Data Engineering: Apache Airflow, Spark, Kafka, Flink, Databricks, Snowflake
- Cloud & DevOps: GCP (GKE, Cloud Run, PubSub), AWS (EKS, Lambda, EC2, S3), Docker, Kubernetes, CI/CD
- Databases: PostgreSQL, MongoDB, Neo4j, Amazon RDS

## Certifications

- Microsoft Azure AI Fundamentals (AI-900)
- Microsoft Azure Fundamentals (AZ-900)
- Microsoft Azure Developer Associate (AZ-204)
- CompTIA Security+ (SY0-701)