

Jignesh Ammineni

(413) 404-2123 | jigneshammineni@gmail.com | <https://www.linkedin.com/in/jignesh-ammineni> | Palo Alto, CA

EDUCATION

University of Massachusetts, Amherst

Amherst, MA

Bachelor of Science in Computer Science

Graduated December 2025

- Chancellor's Award Scholarship Recipient — Dean's List Student
- **Relevant Coursework:** Systems Design; Software Development Life Cycle; Design Patterns; Data Structures and Algorithms; Object Oriented Programming; Data Science; Big Data Analytics; Machine Learning; OS

TECHNICAL SKILLS

Languages: Python, JavaScript, TypeScript, Java, C/C++, HTML & CSS

Frontend & Frameworks: Next.js, React, Playwright, A/B Testing, Beautiful Soup

Backend & APIs: RESTful APIs, Docker, Amazon ECS, FastAPI, Node.js, MongoDB, PostgreSQL, Github Actions

ML/Data: Prompt Engineering, LangChain, RAG, RAGAS, Vertex AI, Finetuning LLMs, Scikit-Learn

EXPERIENCE

AI Agents Researcher

Jan. 2026 - Apr. 2026

- Co-authored 2 research papers for COLM investigating **multi-agent orchestration** patterns and vulnerabilities in the **Agent-to-Agent (A2A) Protocol**, and building safety frameworks for coordination and protocol-level attacks
- Built AI/ML pipelines for multi-agent systems and Language Models, implementing automated dataset curation, ablation frameworks, and benchmark harnesses to identify AI agent coordination, communication patterns, and failure modes

PROJECTS

Agentic AI Engineer — Voice Agent

- Built a real-time voice AI agent that holds natural conversation at 1.8s median latency to act as a thinking partner, asking guiding questions, pressure-testing your reasoning, and outputting structured artifacts
- Engineered a voice pipeline with Pipecat to stream 20ms chunks of audio through Speech-to-Text → LLM → Text-to-Speech, with automatic turn detection and barge-in (mid-response interruption) features
- Architected a tiered memory layer (MemGPT framework) with self-directed async tool calls for recursive summarization and vector-search retrieval, automating context engineering without stalling conversation
- Deployed on a GCP VM with WebRTC transport, enabling packet-loss tolerance and avoiding TCP's head-of-line blocking and retransmission latency

AI/ML Engineer — Insurance Policy RAG Chatbot

- Indexed 1,500+ pages of insurance policy into a retrieval system that turns minutes of document-reading into seconds of Q&A with a chatbot for policyholders, answering at 96% correctness
- Engineered a modular **RAG pipeline** using LangChain, OpenAI, and Pinecone vector database, implementing semantic chunking, hybrid retrieval (cosine sim + BM25 via Reciprocal Rank Fusion) and CoT prompt engineering
- Built an automated evaluation pipeline using **RAGAS framework** and LLM-as-judge against a curated golden dataset of 50+ Q&A pairs, scoring 94% context precision, and 93% faithfulness

Agentic AI Engineer — Multi-Agent Content Generation System

- Built an automated content creator system on Google's **ADK** to produce sports news posts with captioned images
- Orchestrated **specialist** sub-agents (researcher, writer, image generator, editor) with ADK's hierarchical multi-agent patterns, integrating Gemini 2.5 Pro for narrative and OpenAI 4o Image Generation
- Deployed to **Vertex AI's** Agent Engine runtime, with an ADK **evaluation harness** that scores trajectory correctness and factual accuracy via LLM-as-judge

Backend Software Engineer — Distributed Microservices

- Engineered a containerized, distributed Movie Watchlist service's backend with 6 **FastAPI** microservices enforcing **database-per-service** architecture (PostgreSQL + SQLAlchemy), exposing 32 RESTful endpoints
- Implemented an **NGINX API Gateway** with round-robin load balancing across multiple service instances for horizontal scaling to **5x baseline throughput**, and Dockerized all services to standardize integration testing and deployment
- Identified read-heavy services from traffic patterns and implemented Redis caching with the cache-aside pattern, reducing average **response latency** by **40%** on repeated reads