Ryan Yavari

San Francisco, CA (Open to Relocate) | 408-892-7723 | ryanyavari10@gmail.com | linkedin.com/in/ryan-yavari | ryanyavari.com

EDUCATION

Georgia Institute of Technology (Part-Time, Online) Atlanta, GA
Master of Science in Computer Science (Specialization in Artificial Intelligence). GPA: 3.67/4.0 *Expected Dec 2027*

San Jose State University San Jose, CA
Bachelor of Science in Computer Science. GPA: 3.67/4.0 *May 2025*

EXPERIENCE

Software Engineer Intern June 2025 - Jan. 2026
San Jose Mineta International Airport (SJC) *San Jose, CA*

- Engineered and deployed an LLM-powered customer service system with tool-augmented RAG pipeline serving 10,000+ travelers, reducing call volume by 17% and saving $11,000+ in annual support costs, live at flysanjose.com
- Implemented a production-grade RAG pipeline in Python with tool calling using LangChain, OpenAI embeddings, and Qdrant, enabling grounded, context-aware responses across 13,000+ chunked documents ingested from 667 pages of the SJC website
- Built a Redis-based semantic cache with hash-value change detection and automated refresh, reducing LLM inference costs by 25% and improving end-to-end response latency by 66%
- Architected a serverless deployment on Google Cloud Functions and Firebase with retry logic, graceful failure modes, and elastic scaling, maintaining 99.99% uptime across 6+ months in production
- Designed and ran a Python-based model selection evaluation suite across GPT variants, using an LLM-as-Judge on 200+ benchmark Q&A pairs, improving answer accuracy from 75% to 98% and reducing response latency from 5s to under 3s

PROJECTS

Optimal (AI Productivity Agent)| https://useoptimal.app June 2025 - Jan. 2026

- Engineered and launched an AI productivity platform used by 100+ users, built on a supervisor-orchestrated multi-agent system with stateful routing and structured agent handoffs
- Designed human-in-the-loop (HITL) approval gates where agents generate structured proposals for user review before persistence, enabling safe, auditable agentic actions with zero unreviewed writes
- Built offline evaluation benchmarks and regression suites to measure agent task completion, output quality, and planning correctness, cutting prompt regression incidents by 40% across 3 model update cycles
- Built full-stack infrastructure with Java Spring Boot, React Native (Expo), and secure token-based auth, designing a Supabase PostgreSQL schema with 17+ entities using EER modeling and normalization

PrepAgent (AI Interview Research Agent) | https://github.com/ryanyavari/prepagent Sept. 2026 - Nov. 2026

- Built and deployed a full-stack AI interview-prep app (FastAPI, Next.js, Supabase pgvector, AWS EC2, Vercel), live at useprepagent.app, that replaces manual pre-interview research with a sourced 4-section briefing
- Designed a LangGraph fixed-DAG pipeline (parallel company and interviewer agents, then RAG, then synthesis) with per-call timeouts, retries and a hard deadline, so a stalled API call fails fast instead of hanging
- Built a RAGAS eval harness with versioned golden datasets that exposed weak synthesis faithfulness (0.56); grounding synthesis in the job description raised it to 0.99 on 3 synthetic cases (self-graded)
- Integrated Google Calendar OAuth to auto-detect interview events with a prompt-injection-hardened LLM classifier, plus encrypted token storage, rate limits and per-user cost caps, backed by 690 tests in CI

Quizzler (AI Quiz Generator) | https://github.com/ryanyavari/quizzler Mar. 2025, Updated Apr. 2026

- Instrumented a 5-stage RAG pipeline with LangSmith tracing across query expansion, embedding, reranking, and generation, capturing token usage, Cohere relevance scores, and per-span latency to surface retrieval quality and cost regressions
- Architected parallel upload processing to chunk and embed user documents into an ephemeral server-side vector store while concurrently generating structured MCQ quizzes, delivering grounded answers with glass-box citations from user-uploaded PDFs

SKILLS

Languages: Python, SQL, TypeScript, JavaScript, Java, Golang (Go), C, C++, HTML, CSS
AI & Machine Learning: AI Agents, Generative AI, LLMs, Tool Calling, MCP, RAG, Prompt Engineering, Agentic Systems, Workflow Automation, Responsible AI, Anthropic API, OpenAI API, Function Calling, LangChain, LangGraph, LLM Eval, RAGAS, Hallucination Detection, Observability, Vector Databases, Data Preparation
Core Frameworks: FastAPI, Flask, Node.js, Express, React, Next.js, Spring Boot, React Native (Expo)
Databases & Cloud: PostgreSQL, MySQL, Redis, Supabase, pgvector, Qdrant, Pinecone, MongoDB, GCP, AWS, Firebase, Azure
Tools: RESTful APIs, OAuth, Git, Docker, CI/CD, Postman, LangSmith, LlamaIndex, Vercel AI, Linux, Kubernetes, System Design