

Ryan Yavari

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

EDUCATION

Georgia Institute of Technology

Master of Science in Computer Science (Specialization in Artificial Intelligence). GPA: 3.67/4.0

Atlanta, GA

Expected May 2027

San Jose State University

Bachelor of Science in Computer Science. GPA: 3.67/4.0

San Jose, CA

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 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 domain
- 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 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

PrepAgent (AI Interview Research Agent) | <https://github.com/ryanyavari/prepagent>

June 2026 - Present

- Architecting a MCP-native multi-agent system using LangGraph supervisor graph with 4 specialized agents, integrating Gmail, Google Calendar, and Google Drive via Google OAuth to autonomously detect interview invites and trigger a personalized research pipeline without manual user input
- Engineering a RAG pipeline with Pinecone vector store, Cohere Rerank, and OpenAI text-embedding-3-small to retrieve and rerank context across heterogeneous sources (resume, past briefings, company docs), with RAGAS eval suite tracking answer relevance, faithfulness, and context precision across 20 ground truth briefings
- Engineering a system to automate pre-interview research that currently takes 45 minutes of manual review across resume, company docs, and past briefings, targeting sub-2-minute end-to-end generation

Optimal (AI Productivity Agent) | <https://useoptimal.app>

June 2025 - Jan. 2026

- Engineered an AI productivity platform using a supervisor-orchestrated multi-agent system with stateful routing and structured agent handoffs, reducing malformed outputs by 60%
- 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 Spring Boot, React Native (Expo), and secure token-based auth, designing a Supabase PostgreSQL schema with 17+ entities using EER modeling and normalization

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 and text

SKILLS

Languages: Python, TypeScript, SQL, JavaScript, Java, HTML, CSS

AI & Machine Learning: RAG, Agentic Systems, Tool Calling, LangChain, LangGraph, MCP, OpenAI API, Anthropic API, Vector Databases, LLM Eval, RAGAS, Hallucination Detection, Observability, Function Calling, Prompt Engineering

Core Frameworks: FastAPI, Flask, Spring Boot, Node.js, Express, React, Next.js, React Native (Expo)

Databases & Cloud: PostgreSQL, Redis, GCP, AWS, Firebase, MongoDB, MySQL, Supabase, pgvector, Qdrant, Pinecone

Tools: Linux, Docker, Kubernetes, Git, CI/CD, RESTful APIs, System Design, Postman, LangSmith, LlamaIndex, Vercel AI