

Rahul Muralidhara Koushik

Email: rkoushik@asu.edu

GitHub: github.com/Fluff18 | Portfolio: fluff18.github.io

San Francisco Bay Area

Mobile: +1-480-791-6597

LinkedIn: linkedin.com/in/rahulkoushik

AI Engineer | LLM Agents, Evals, Inference. Builds generative AI (GenAI) and large language model (LLM) agents end to end: the user problem, the agent loop, the memory layer and the evals that decide whether an agent is ready to ship. Three HPE hackathon wins, two selected for Fellow Review and BU-wide rollout; patent under way; HPE Early Career Technologist.

EXPERIENCE

Hewlett Packard Enterprise, Storage AI Team

AI Engineer

San Jose, US

Feb 2026 – Present

- Architected the production **multi-agent LLM harness** with **guardrails** that every team agent runs on; its **Model Context Protocol (MCP)** tool server handles **6,300+ tool calls a week** across HPE Storage.
- Built the **memory layer** for persistent agents: episodic memory, **embeddings and vector search**, and a **context graph** over **13,005** mined bugs and **5,600** grounded analyses.
- Cut agent latency and cost per run with **inference optimization**: **prompt caching** and **KV-cache reuse** across agent steps, **cache-coherent** context updates, and **prefill/decode-aware** serving of **Qwen** models (27B, 35B) on **vLLM**.
- Took the **agent evaluation platform** from idea to **three production deployments in four weeks**; rubric-based evals, **LLM-as-a-judge** scoring and a **tracing dashboard** for every agent run from the **release gate**.
- Shipped **agentic alert-triage and root-cause-analysis (RCA)** systems to production, with **self-improving loops** that learn from their own failures.

Hewlett Packard Enterprise, Storage

Cloud Software Engineer

San Jose, US

Jun 2023 – Feb 2026

- Built the **observability platform** for **10K customers**: cross-stack root-cause analysis cut resolution time from **30 to 3 minutes** and raised first-time fix success from **55% to 85%**.
- Shipped **6 Peer Motion features** in Go across four storage array families, including the Nimble-to-Alletra migration path.
- Built a **documentation agent** that updates docs on every pull request, saving **5 engineer-hours a week** per team.

The Luminosity Lab, Arizona State University

Research Aide

Tempe, US

Feb 2022 – May 2023

- Led the iOS team on the TouchPoint apps; a telehealth flow saved doctors **15 minutes** per visit.

Aruba, Hewlett Packard Enterprise

Software Engineering Intern

Tempe, US

May 2022 – Aug 2022

- Built a test-automation tool that configures network topologies, saving an estimated **150k engineering hours a year**.

Webex by Cisco

Technical Undergraduate Intern

Bengaluru, India

Jan 2021 – Jul 2021

- Built call and meeting scheduling APIs for the Webex iOS SDK v3 and the Kitchen Sink sample app.

SKILLS

- **Generative AI (GenAI)**: LLM agents, agentic workflows, MCP and tool calling, evals and LLM-as-a-judge, Retrieval-Augmented Generation (RAG), embeddings and vector search (pgvector), guardrails, multi-agent orchestration, prompt and context engineering, memory, tracing, LangGraph, vLLM, Qwen, llama.cpp
- **Languages**: Python, Go, TypeScript, JavaScript, Swift, Kotlin, C++, SQL
- **Infrastructure**: Distributed systems, Kubernetes, Helm, Docker, AWS, Kafka, Spark, Postgres, Prometheus, Grafana, gRPC, FastAPI, React, Next.js

SELECTED PROJECTS

- **Zuo, personal AI harness (2026)**: Agent teams, job queue and budgets run from my phone across three machines.
- **Catan Arena, self-learning agents (2026)**: Agents graduate only when a statistical pass bar says the learning is real.
- **Zzz, iOS and Android (2026)**: Gamified goal tracker, live on the App Store. SwiftUI, Kotlin, Supabase.

EDUCATION AND PUBLICATIONS

Arizona State University

Master of Science, Computer Science

Tempe, US

Aug 2021 – May 2023

PES University

Bachelor of Technology, Computer Science and Engineering

Bengaluru, India

Aug 2017 – Jul 2021

Publications: ViStA: Video Streaming and Analytics Benchmark, IEEE ISPASS 2021; Generation of True Random Numbers using Entropy Sources in Portable Computers, IEEE CONECCT 2021; PipeGAN, a pipelined parallel GAN, Springer LNNS vol. 171, 2020.