Rahul Muralidhara Koushik

San Francisco Bay Area
Email: rkoushik@asu.edu
Mobile: +1-480-791-6597
GitHub: github.com/Fluff18 | Portfolio: fluff18.github.io
LinkedIn: linkedin.com/in/rahulkoushik

AI Engineer and AI Product Builder | LLM Agents, Evals, Inference. I build generative AI (GenAI) and large language model (LLM) products end to end: the user problem, the agent loop, the memory layer and the evals that decide whether an agent is ready to ship. Interested in AI engineering, AI product and physical AI.

Product Highlights

Defined the ship bar for AI agents at HPE Storage: worked with storage experts to turn "is this agent good?" into rubric-based evals and a release gate, then took the platform from idea to three production deployments in four weeks.

Built for adoption, not mandate: the MCP tool server became the shared path every team agent and the experts' Copilot use, at 6,300+ tool calls a week.

Chose honesty over a win: reported a zero-lift result on self-learning agents after finding data leakage, which made the team trust the evaluation platform more.

Owned a consumer app end to end: scoped, designed, built and launched **Zzz** on the App Store, from the goal-and-reward model to the privacy site.

Product decisions with clients: at the Luminosity Lab, led the iOS team and shaped the client's mobile roadmap; the telehealth flow saved doctors 15 minutes a visit.

Zero-to-one under constraints: won three HPE hackathons, two taken into BU-wide production through Fellow Review; Best Project at a lean-startup program across 34 teams.

Experience

Hewlett Packard Enterprise, Storage AI Team — San Jose, US
AI Engineer — *Feb 2026 – Present*

- Built the production **multi-agent LLM harness** with **guardrails** that every team agent runs on. I also own its **Model Context Protocol (MCP)** server and all of its tools, used across HPE Storage software engineering and by the storage experts' Copilot: **6,300+ tool calls a week**.
- Engineered the **memory layer** for persistent agents: episodic memory across runs, **embeddings and vector search** on pgvector, and a **context graph** linking **13,005** mined bugs and **5,600** grounded analyses by component, symptom and fix, built in nine days.
- Cut agent latency **45%** with **cache-coherent** context updates, alongside **prompt caching**, **KV-cache reuse** across agent steps and **prefill/decode-aware** serving of **Qwen** models (27B, 35B) on **vLLM**.
- Scaled the **agent evaluation platform** from idea to **three production deployments in four weeks**: rubric-based evals scored against expert feedback, **LLM-as-a-judge**, and a **tracing dashboard** showing every step, tool call and failure of each agent run. It is the **release gate** every agent must pass.
- Shipped **agentic alert-triage and root-cause-analysis (RCA)** systems to production, with **self-improving loops**: a reflector agent reviews every run to check the agent took the right path, and subject-matter experts confirm the corrections.

Hewlett Packard Enterprise, Storage — San Jose, US
Cloud Software Engineer — *Jun 2023 – Feb 2026*

- Developed the **observability platform** used by the internal teams supporting **30K devices** and **10K customers**: log-level root-cause analysis across hardware, OS, backend, UI and cloud cut resolution time from **30 to 3 minutes**; first-attempt array setup success rose from **55% to 85%**.
- Delivered **6 full-stack enterprise data-migration features** (HPE Peer Motion) in Go and JavaScript across four generations of storage arrays (3PAR, Primera, Alletra, Nimble), over Fibre Channel and iSCSI: Remote Copy support, the Nimble-to-Alletra iSCSI migration path, and domain, multi-link and VLAN handling. **15 tagged builds**, **47 defects** fixed across **5 release lines**.
- Launched a **documentation agent** over a vector knowledge base that updates docs and API references on every pull request, saving about **5 engineer-hours a week per team**.
- **2024 HPE hackathon winner** with an end-to-end RAG chatbot over PDFs; named **HPE Early Career Technologist**.

The Luminosity Lab, Arizona State University — Tempe, US
Research Aide — *Feb 2022 – May 2023*

- Led the iOS team revamping the TouchPoint iOS and Android apps and helped the client make mobile product decisions.
- Used HealthKit heart-rate-variability and sensor data to classify users into six stress personalities.
- Designed a telehealth flow on the Zoom Meeting SDK, with Phoenix Children's Hospital, that lets doctors start remote visits and drive the patient's Android TV, saving **15 minutes** per visit.

Aruba, Hewlett Packard Enterprise — Tempe, US
Software Engineering Intern — *May 2022 – Aug 2022*

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

– Presented the product to panels of senior engineers and cross-functional representatives.

Webex by Cisco Bengaluru, India
Technical Undergraduate Intern *Jan 2021 – Jul 2021*

– Added call and meeting scheduling APIs to the Webex iOS SDK v3.
– Wrote Kitchen Sink, the sample iOS app that demonstrates the SDK's APIs.

Skills

• **Generative AI (GenAI)**: LLM agents, agentic workflows, multi-agent orchestration, MCP and tool calling, evals and LLM-as-a-judge, Retrieval-Augmented Generation (RAG), embeddings and vector search (pgvector), guardrails, prompt and context engineering, memory, tracing, LangGraph, vLLM, Qwen, llama.cpp, Ollama, Whisper

• **Languages**: Python, Go, TypeScript, JavaScript, Swift, Kotlin, C++, SQL, Shell

• **Infrastructure**: Distributed systems, Kubernetes, Helm, Docker, AWS (Lambda, SQS, EC2, S3, DynamoDB), Kafka, Spark, Trino, Postgres, Supabase, Firebase, Prometheus, Grafana, Humio, gRPC, REST, FastAPI, Node.js, Jenkins CI/CD, Git, Linux

• **Apps and UI**: React, Next.js, Tailwind, SwiftUI, UIKit, HealthKit, Android, Playwright

• **Product**: Problem scoping with users, success metrics and pass bars, zero-to-one delivery, cross-functional presentations

Projects

• **Zuo, personal AI harness (2026)**: Job queue, budgets, project board and agent teams controlled from my phone; free NVIDIA endpoints, a PC GPU on vLLM and a laptop on llama.cpp as workers, Claude as planner. Voice capture with silence gating and Whisper; an iOS client; a UI agent team that turns design posts into web sections. Python, FastAPI, LangGraph, Docker.

• **Catan Arena, self-learning agents (2026)**: One learning agent plays full Catan against three frozen copies of itself; it graduates only when the 95% Wilson interval of its win rate clears the 25% null. Actor, watcher and consolidator loop with persistent memory.

• **AI Tutor (2026)**: Adaptive learning app for AI engineering: a level path, quizzes and a swipeable feed built from Reddit, Hacker News, arXiv and YouTube; Claude with an automatic NVIDIA Nemotron fallback.

• **Zzz, iOS and Android (2026)**: Gamified goal tracker with quests, points and rewards, live on the App Store. SwiftUI, Kotlin, Supabase. App Store

• **Drone inspection simulation (2026)**: PX4 and Gazebo inspection drone with an agent capped to Raspberry Pi 5 resources: YOLOv8 plus a small vision-language model; llama.cpp 18x faster than Ollama on the same limits.

• **musaic (2026)**: Automated lo-fi music channel: Stable Audio, GPT-4 metadata, DALL-E and Imagen thumbnails, Veo backgrounds, FFmpeg encode and YouTube upload. GitHub

• **Quantum slot machine (2025)**: Randomness from real qubits on IBM Quantum hardware; Qiskit and FastAPI backend, React frontend. GitHub

• **Earlier (2020 to 2022)**: RideShare, a fault-tolerant database service on AWS with RabbitMQ and Zookeeper; face recognition on AWS Lambda and SQS (93% accuracy) and on autoscaled EC2 (100 concurrent requests); handwriting recognition split across phones (30% faster); Fetch Rewards recipe app in Swift.

Recognition and Leadership

• **HPE**: Early Career Technologist (2024).

• **PES Center for Innovation and Entrepreneurship**: Led a team of six from idea to prototype of a smart attendance wristband; Best Project across 34 teams.

• **Stanford Continuing Studies (Fall 2026)**: Funding a New Enterprise; Building Interpersonal Skills (COM 19), the T-group workshop modelled on Stanford GSB's Interpersonal Dynamics.

Publications

N. Raju, **R. M. Koushik**, H. Om, S. Kalambur. *ViStA: Video Streaming and Analytics Benchmark.* IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS), 2021. doi:10.1109/ISPASS51385.2021.00019

R. M. Koushik, A. Perichiappan, H. Om, A. Banerji, S. Eswaran, P. Honnavalli. *Generation of True Random Numbers using Entropy Sources Present within Portable Computers.* IEEE CONECCT, 2021. doi:10.1109/CONECCT52877.2021.9622734

R. Chandan, N. Pentapati, **R. M. Koushik**, R. Nagpal. *A High Performance Pipelined Parallel Generative Adversarial Network (PipeGAN).* Lecture Notes in Networks and Systems, Springer, vol. 171, pp. 769–777, 2021. doi:10.1007/978-981-33-4543-0_81

Education

M.S. Computer Science, Arizona State University, Tempe. GPA 3.59 Aug 2021 – May 2023

B.Tech Computer Science and Engineering, PES University, Bengaluru. GPA 8.2 Aug 2017 – Jul 2021