

RAHUL KUNDAPUR RAMESH

Raleigh, NC | (602) 394-3489 | ✉ rahulkundapur@gmail.com | in [in/rahul-kundapur](https://www.linkedin.com/in/rahul-kundapur) | 🌐 [rahul-k-r](https://rahul-k-r.github.io)

Work Experience

Traction Layer AI

Nov. 2025 – Present

Founding Engineer

Raleigh, NC

- Engineered a low-latency SSE streaming engine in Go over a unified dispatch seam, with TTFT/TPOT measured on the path that serves traffic; validated against a sub-2ms p50 framing budget via live loopback benchmarks.
- Built a multi-provider API translation layer (CanonReq) in Go normalizing OpenAI and Anthropic payload dialects – image, tool_use, tool_result – across four backends while preserving raw dispatch bytes.
- Engineered dynamic tuning logic in Python for an SGLang inference harness, adjusting `prefill_max_requests` and `chunked_prefill_size` against p95/p99 tail SLOs instead of window means; a 5% throughput gate auto-reverts unproductive moves, catching a runaway that cut throughput from 9,225 to 8,662 tok/s.
- Resolved data plane concurrency edge cases in Go via atomic state distribution using temp-file rename strategies and OS file-lock retry logic, eliminating partial reads under concurrent polling.
- Authored Kubernetes security posture via Helm: default-deny egress NetworkPolicies for in-VPC inference, with the VPC CNI addon pinned so datapath rolls cannot silently disable enforcement.
- Built a post-commit secrets scanner flagging API keys and credential patterns in staged files; closed a cross-tenant IDOR on unscoped querysets and enforced single-active-row invariants with partial unique indexes.
- Migrated the platform from Azure to AWS: 3-tier VPC design, zero-data-loss RDS cutover, and GitHub Actions with OIDC keyless federation removing long-lived credentials from CI.
- Architected a content-agnostic agentic editing framework on LangGraph with dependency injection; targeted-operations editing replaced full-document regeneration, cutting token consumption by ~60%.

Cognizant Technology Solutions

Sept. 2021 – June 2023

Jr. Software Engineer

Bengaluru, India

- Engineered a Python/Pandas security data pipeline to analyze an 1,800-item vulnerability backlog, enabling automated risk prioritization that reduced critical security threats by 25%.
- Led incident response for production outages on AWS, investigating cloud logs, restoring SLA compliance, and authoring post-mortems.
- Re-architected core authentication services in .NET/C#, replacing hardcoded secrets with secure service accounts to eliminate critical security vulnerabilities.

Personal Projects

Portfolio RAG Chat | *Next.js, TypeScript, Gemini, Upstash Redis/Vector*

2026

- Built a streaming RAG chatbot against a 20-request/day free tier: a semantic answer cache serves repeat questions in ~40ms, and quota exhaustion degrades to embedding-matched canned answers before a 503 with upstream-derived Retry-After.

Serverless Video Analysis Pipeline on AWS | *AWS Lambda, Python, S3, Docker, OpenCV*

2024

- Engineered an event-driven serverless pipeline achieving sub-270s processing latency for 100 concurrent streams; optimized Docker multi-stage builds to cut image footprint by 60%.

Education

Arizona State University

Tempe, AZ

Master of Science in Computer Science

May 2025

- Coursework:** Software Security, Cloud Computing, Artificial Intelligence, Data Processing at Scale

Dr. Ambedkar Institute of Technology

Bengaluru, India

Bachelor of Engineering in Computer Science

May 2021

- Coursework:** DBMS, Operating Systems, Neural Networks, Cloud Computing

Technical Skills

Languages: Go, Python, TypeScript, JavaScript, C#, C++, Java, SQL

Backend & Inferencing: Go Microservices, SSE Streaming (TTFT/TPOT), SGLang, vLLM, REST APIs, LangGraph, Django REST Framework

Security: Secret Scanning, Vulnerability Triage, Kubernetes NetworkPolicies, AWS OIDC Federation, OAuth, JWT

DevOps & Cloud: AWS (EKS, ECS, Lambda, S3, RDS, Bedrock), Azure, Docker, Kubernetes, Helm, Terraform, GitHub Actions, CI/CD