

Aditya Karnam Gururaj Rao

Senior Software Engineer

GREATER ST. LOUIS • adityakarnam.grao@gmail.com • (314) 696-4520
adityakarnam.com • github.com/adityak74 • [linkedin.com/in/adityakarnamgrao](https://www.linkedin.com/in/adityakarnamgrao) • [Google Scholar](#)

PROFESSIONAL SUMMARY

Senior Software Engineer with 12+ years across AI/ML, distributed systems, and cloud infrastructure. Tech Lead at Zefr, architecting **ZAiN**, a production AI agent platform spanning MCP servers, tool execution, evals, permissioning, and observability, serving 500+ internal and 10,000+ external users. Delivered \$350K+ in direct business value, including Meta measurement systems with 83% performance gains and vector search infrastructure over 100M records. Published AI researcher (IEEE, Frontiers, arXiv) and open-source maintainer with contributions accepted by Anthropic. Previously founded and scaled a SaaS platform serving 100+ clients.

EMPLOYMENT HISTORY

ZEFR Jan 2022 – Present
Greater St. Louis, Remote, promoted from Software Engineer to Senior Software Engineer across four levels

Senior Software Engineer, Tech Lead of Customer Experience (*Dec 2025 – Present*)

- ◆ Architect **ZAiN**, Zefr's production AI agent platform, spanning agents, MCP servers, tool execution, evals, agent gateway, permissioning, session compaction, and chat interface, including MCP apps for Claude Code, Codex, and Microsoft Copilot 365; serving 500+ internal and 10,000+ external users at 30K+ requests/month in pilot.
- ◆ Build agent observability infrastructure and dashboards tracking top agent users, tool-call volume, and production behavior from agent logs, enabling reliability evaluation of live AI workflows.
- ◆ Lead distributed system design spanning Kafka-based campaign processing, Airflow orchestration, search infrastructure, and API gateway architecture (Apigee/n8n), delivering customer-facing and internal platform systems across AWS and GCP.
- ◆ Author 20+ RFCs and lead cross-team design reviews across 4 teams, mentoring 5+ engineers and setting long-term technical direction, engineering standards, and platform reliability.

Software Engineer to Software Engineer III (*Jan 2022 – Dec 2025*)

- ◆ Delivered \$350K+ in measurable business value across measurement, search, and reporting systems, including two company hackathon wins shipped to production.
- ◆ Built Meta measurement systems (Python, Kafka, Redshift) achieving 83% performance gains.
- ◆ Designed and scaled vector search infrastructure spanning 100M+ records (Elasticsearch).
- ◆ Developed, optimized, and maintained the inventory availability endpoint for digital properties, processing data across multiple Kafka topics on AWS (ECS, Lambda, Firehose, S3).
- ◆ Built a Redshift metrics pipeline publishing custom metrics to AWS CloudWatch and Datadog.

DIGITAL SYSTEMS ARCHITECT Jun 2021 – Jan 2022
ePlata, Greater St. Louis

- ◆ Built NFT platform on Cardano (Plutus Smart Contracts).
- ◆ Architected KYC Remittance API integrating Western Union APIs.
- ◆ Spearheaded IPC-V remittance platform generating \$3M transaction volume.
- ◆ Delivered full-stack engineering for 8+ white-label fintech clients (React, TypeScript, GraphQL, Docker).

FULL STACK ENGINEER Feb 2020 – May 2021
ePlata, Greater St. Louis

- ◆ Migrated AWS EC2 workloads to ECS to optimize resource utilization and reduce costs.
- ◆ Mentored interns and built GraphQL API testing templates.
- ◆ Rebuilt logging service with support for Sentry, Logzio + PII sanitization.

SOFTWARE ENGINEER Jun 2018 – Jan 2020
ePlata, Greater St. Louis

- ◆ Integrated Splunk ML Toolkit for Payment Fraud System.
- ◆ Improved build pipelines with AWS CodeBuild (75% time reduction).
- ◆ Built native contacts integration (Swift, Kotlin, TypeScript) increasing load performance by 60%.
- ◆ Created CI/CD pipelines with Docker, Jenkins, CircleCI, AWS CodeDeploy.

GRADUATE RESEARCH ASSISTANT Aug 2017 – Jun 2018
University of Missouri–St. Louis, Greater St. Louis

- ◆ Thesis: *Efficient Reduced-Bias Genetic Algorithm for Generic Community Detection Objectives*, a novel genetic algorithm

framework for clustering networks with 5M+ nodes.

- ◆ Collaborated with Oak Ridge National Lab, using NVIDIA Titan cluster computing resources on an Alzheimer's gene-expression network dataset.
- ◆ Reduced memory usage with a 3-dimensional bit-array container.

FOUNDER & CTO

Aug 2014 – Aug 2016

eventzé

- ◆ Founded a SaaS platform for managing large college events covering registrations, merchandise, finance automation, and digital branding, partnering with 100+ college fests.
- ◆ Built and led a team of 10+ developers to scale the platform (PHP/CodeIgniter, ReactJS, REST APIs, payment gateway integration, DevOps automation).

SELECTED PROJECTS & OPEN SOURCE

- ◆ **Call to Think**: personal essay blog on technology, AI, and society, built with Astro and deployed on Cloudflare Workers.
- ◆ **quecto**: a tiny OpenAI-compatible AI harness and coding agent in Rust with a ~1.3 MB core and ~3.5 MB agent binary.
- ◆ **embenx**: a universal embedding retrieval toolkit and agentic memory layer with a unified API for 15+ vector backends and MCP support for Claude and autonomous agents (`pip install embenx`).
- ◆ **locobench**: local coding model benchmarking, a Harbor and Ollama setup for running local agentic benchmarks with a leaderboard pipeline.
- ◆ **mcp-scholarly**: academic retrieval with MCP for scholarly search from AI assistants (185+ GitHub stars).
- ◆ **opencontext**: interoperable AI chat contexts across assistants.
- ◆ **medfit-llm**: applied model fine-tuning in healthcare.
- ◆ **google-drive-upload-git-action**: developer productivity automation as a reusable GitHub Action (108+ GitHub stars).

EDUCATION

MS, Global Business and Blockchain , University of the Cumberland	2021 – 2022
MS, Computer Science , University of Missouri–St. Louis	2016 – 2018
B.Tech, Computer Science , Vellore Institute of Technology	2012 – 2016

PUBLICATIONS & RESEARCH

- ◆ *MEDFIT-LLM: Medical Enhancements through Domain-Focused Fine Tuning of Small Language Models*. IEEE, 2025.
- ◆ *CAG: Chunked Augmented Generation for Google Chrome's Built-in Gemini Nano*. arXiv:2412.18708, 2024.
- ◆ *AI-Enhanced Neural Network Framework for Cardiac Arrest Prognostics*. Frontiers in Health Informatics, 2024.
- ◆ *Automated Analysis of Glucose Response Patterns in Type 1 Diabetes Using Machine Learning and Computer Vision*. International Journal of Advanced Computer Science & Applications, 2025.
- ◆ *Neural Network-Enabled Diagnostic Framework for Correlating Ayurvedic Breathwork Protocols with Adaptive Skin Microbiome Dynamics*. Mathematical Modelling of Engineering Problems, 2025.
- ◆ *The Harness Engineer*. Book, 2025.
- ◆ *Quantifying Adaptive Governance Success: A Mixed-Methods Analysis of 500 Organizations*. 2025.
- ◆ *Efficient Reduced-Bias Genetic Algorithm (ERBGA) for Generic Community Detection Objectives*. MS thesis, University of Missouri–St. Louis, 2018.
- ◆ *Transmission in Benes Network*. International Journal of Pure and Applied Mathematics, 2016.