

Adithyan K

AI Research Engineer | Agent Systems, Evaluation & LLM Infrastructure

adithyn.krshna@gmail.com | +91 7034046267 | adithyn.dev | [GitHub](https://github.com/adithyn) | [LinkedIn](https://www.linkedin.com/in/adithyan-k) | [Google Scholar](https://scholar.google.com/citations?user=adithyan)

EDUCATION

Manipal Institute of Technology

Bengaluru, India

B.Tech in Computer Science and Engineering (AI), CGPA: 8.41

Sep. 2022 – Jun. 2026

- **Relevant Coursework:** Machine Learning, Deep Learning, Natural Language Processing, Generative AI, Reinforcement Learning, Data Structures & Algorithms

PUBLICATIONS

Scouting: Cost-Aware Routing of Coding Agents by Scouting the Repository First

[arXiv:2608.04804](https://arxiv.org/abs/2608.04804)

SuperAGI Research, Aug 2026

Built the system and ran the experiments: training pipeline, the verification gate that strips unsupported reproduction claims, the cost-aware router, and the evaluation harness.

Code Isn't Memory: A Structural Codebase Index Inside a Coding Agent

[arXiv:2606.22417](https://arxiv.org/abs/2606.22417)

SuperAGI Research, Jun 2026

Built the evaluation platform and ran the experiments, including the leak-audit protocol and localization metric.

EXPERIENCE

AI Research Engineer

Jul 2026 – Present

SuperAGI

Bengaluru, India

- **SuperScout:** Trained a 7B repository-search model end to end and built the cost-aware routing system around it: contamination-controlled curation of a 19.9K-example training corpus, a custom sample-packing pipeline, LoRA fine-tuning of Qwen2.5-Coder, a verification gate that replays and strips unsupported reproduction claims, and a hidden-state router. Tech: PyTorch, PEFT, GRPO, vLLM, Modal. **Matches the best frontier model at about a fifth the cost per solve** on the full SWE-bench Pro Python census; \$86 end-to-end training run.
- **Agent Evaluation:** Built the evaluation harness benchmarking **four frontier models head to head** across providers and coding benchmarks, with matched-cost protocols, sandboxed per-episode execution, and an LLM-usage proxy as the cost source of truth. Tech: Python, Modal, vLLM, SQLite. **3,000+ agentic eval episodes** run across three compute planes.
- **SuperCoder Evals:** Built the evaluation platform behind *Code Isn't Memory*: job orchestration, three-vector leak auditing, fail-closed preflight identity gates, and a localization metric designed not to favour our own system. **819 agent runs** over a 3-arm × 3-seed ablation across 31 repositories and 173 per-commit indexes.

AI Engineer, Intern

Jan 2026 – Jul 2026

- **Vibe Coder:** Built the execution and deployment infrastructure for SuperAGI's production AI app builder: Modal sandboxes with build/serve/proxy tooling that runs LLM-generated full-stack apps and streams a live preview from a prompt, plus a static deploy pipeline on S3 + CloudFront using edge functions for per-app subdomain routing off one distribution. Tech: Go, Modal, AWS S3/CloudFront, Redis, asynq. **450+ apps generated and served.**
- **Digital Employees:** Built the backend for autonomous voice + avatar agents that join client meetings to run sales demos: the agent loop, the STT pipeline over a managed WebSocket, and segment-wise slide control via LLM structured outputs with graceful drain and session rehydration. Tech: Go, LiveKit, Deepgram, ElevenLabs, cgo. **Used across 4,000+ meetings** in production.
- **SuperCoder:** Built the Rust client for the context engine in *SuperCoder*, SuperAGI's open-source local-first coding agent: repository parsing and structure-aware chunking across **10 languages**, with a resumable content-addressed sync protocol. Tech: Rust, tree-sitter.
- **Common Stack:** Go, Python, Rust, PostgreSQL, Redis, Modal, Docker, OpenTelemetry, Bazel.

Co-Founder & Technical Lead

Jul 2025 – Present

Lifie, voice and chat AI for businesses

Remote

- Co-founded a voice + chat AI platform: **REACH** (outbound voice engagement) and **ASSIST** (inbound chat support and sales workflows), and own the architecture end to end.
- Built the entire AI + backend layer of **ASSIST**, a multi-tenant **Mastra**/TypeScript agent platform live in production: per-business RAG on **Qdrant** + **FastEmbed**, multi-pass guardrail validation, tool-calling, and configurable workflows over the **WhatsApp Cloud API**; built **Lifie Hub**, an agent-discovery layer with vector search and **A2A** + **MCP** for cross-platform agent interoperability.

AI Product Development Intern

June 2025 – July 2025

QpiAI

Bengaluru, India

- Built and shipped a complete sales support agent on QpiAI's Agent Hive platform: automated lead research, enrichment, and personalized outreach generation, replacing a previously manual workflow.
- Integrated LLM APIs and internal SLMs for dynamic lead scoring, enrichment, and content generation across the pipeline.

PROJECTS

cc-vault | *Go, Bubble Tea, Lip Gloss*

[GitHub](#)

- Built a terminal UI in **Go** for browsing and managing **Claude Code** sessions: a three-panel layout with full-text search across every conversation, markdown-rendered preview, rename/delete/bulk operations, export to markdown, and resume-in-place; reads local JSONL transcripts directly with optimized partial reads for instant loading.

Metafog.io | *Next.js, Docker, Nginx, Ollama, LlamaIndex*

[Live Link](#)

- Built the full stack of a decentralized AI search engine: a crawl → retrieve → LLM-summarize (**LlamaIndex**) pipeline served by a network of **external provider nodes** contributing GPU inference (**Ollama**) and bandwidth; owned frontend, backend, Docker orchestration, and Nginx load balancing across provider nodes.

ArXiv Research Q&A | *Unsloth, QLoRA, Hugging Face, LangChain, Qdrant, Docker*

[GitHub](#)

- Fine-tuned **Qwen 8B & 14B** with **QLoRA** + **Unsloth** on a domain-specific research corpus and built the end-to-end pipeline: PDF ingestion, chunking, embedding, and **Qdrant** + **LangChain** retrieval-grounded inference to reduce out-of-context hallucinations.

SKILLS

Core: LLM Fine-tuning, Agent Evaluation & Benchmarking, Agent Infrastructure, RAG Pipelines, Backend Architecture

Languages: Python, TypeScript, Go, Rust, JavaScript, SQL

AI/ML: LoRA/QLoRA (Unsloth, PEFT), vLLM, PyTorch, Hugging Face, Multi-Agent Systems, Mastra, LangChain, Context Engineering, Structured Outputs, MCP, A2A, Voice AI (LiveKit), Vector Search (Qdrant, FastEmbed)

Infrastructure: Modal, Docker, AWS (S3, CloudFront), GCP, Nginx, FastAPI, GitHub Actions, PostgreSQL, Qdrant, ChromaDB, tree-sitter