

KEDAR SATHE

AI Infrastructure & ML Systems Engineer | Distributed Training · LLM Inference · Backend Systems
kedarsathe49@gmail.com | Pune, India | wtfkedar.me | github.com/kedar49 | linkedin.com/in/kedar49 | [LeetCode/wtfkedar](https://LeetCode.com/wtfkedar)

SUMMARY

Engineer working between the trained model and the production system: distributed training, inference serving, retrieval pipelines, and backend services. Built a 3D-parallel RLHF stack with custom Triton kernels from scratch; Mermaid.js maintainer with merged work in vLLM, Weaviate, and DeepMind AndroidEnv.

EXPERIENCE

ML / Software Engineer Jun 2026 – Present
Autonex AI 360 Remote · Mumbai, India

- Architected DealBridge from scratch in Go: a microservices platform unifying the GTM-to-engineering lifecycle across teams
 - Owned it end-to-end: service boundaries, auth and RBAC, database and schema design, APIs, and workflows; drove decisions on data ownership, query patterns, and horizontal scalability
 - Reworked the PM Portal for 300–400+ daily concurrent users, optimizing large payloads, queries on growing datasets, and traffic handling
 - Fine-tuned domain LLMs with PEFT (LoRA, QLoRA); containerized inference and benchmarked latency across hardware targets
- Go · Microservices · PostgreSQL · RBAC · Docker · PyTorch · PEFT · System Design*

Software / ML Engineering Intern Jan 2026 – May 2026
Unique Schools App (Irish Startup) Pune, India

- Designed AI workflow infrastructure across mobile and web: request IDs, query routing, task scheduling, workflow state, resolver assignment
 - Built a NotebookLM-style internal RAG platform with page-level indexing over BRDs, POCs, and technical docs, optimized for retrieval quality, grounding, and traceability back to source pages
 - Migrated a production platform from Java + PHP to React Native + Node.js and hardened CI/CD for automated build, test, and deploy
- Python · LangGraph · RAG · pgvector · Node.js · React Native · CI/CD*

Maintainer & Open-Source Contributor 2024 – Present
Mermaid.js · vLLM · Weaviate · Google DeepMind AndroidEnv · Google Quiche Remote

- Active maintainer of Mermaid.js (90K+ stars, embedded in GitHub, GitLab, Notion): PR review and triage; fixed 8+ critical parser bugs
- Merged upstream fixes in vLLM (LLM serving), Weaviate (nested blobHash traversal in recursive SHA-256 hashing, Go), DeepMind AndroidEnv (RL infra), and Google Quiche (HTTP/3 null-pointer crash on 1xx interim responses)

PROJECTS

Small-LLM-RLHF github.com/kedar49/small-llm-rlhf
Distributed RLHF Training from Scratch · PyTorch · Triton · CUDA

- 3D-parallel RLHF training for small LMs on raw PyTorch + Triton, no framework: hand-rolled tensor, pipeline (1F1B), and data parallelism with ZeRO-1 optimizer-state sharding
- Wrote custom Triton FlashAttention-2 kernels and a PagedAttention inference engine with its own KV-cache block manager
- Async PPO loop overlapping rollout generation on one GPU set with backpropagation on another

GroundedRAG groundedrag.vercel.app
Local Multi-Agent RAG · LangGraph · Ollama · PostgreSQL · pgvector · Python

- Built a fully local multi-agent RAG system with hybrid BM25 + dense retrieval fused via RRF, cross-encoder reranking, and HyDE query expansion
- pgvector HNSW indexing for sub-1s p90 latency, plus a refusal path when retrieval is too weak to ground an answer

Cybur [Write-up on Medium](#)
Linux Kernel Module · C · Kernel-Space Networking · Security

- Kernel module that intercepts ARP traffic and cross-checks a DHCP snooping table to stop ARP poisoning attacks

SKILLS

Languages: Python, Go, C/C++, TypeScript/JavaScript, SQL, Bash
Training & Kernels: PyTorch, Triton, CUDA, 3D Parallelism (TP/PP/DP), ZeRO, FlashAttention, RLHF/PPO, PEFT (LoRA, QLoRA), Unsloth, Hugging Face
Inference & Serving: vLLM, PagedAttention, Triton Inference Server, FastAPI, Docker, Kubernetes
Retrieval & Data: LangGraph, LangChain, LlamaIndex, RAG (hybrid retrieval, RRF, reranking), pgvector (HNSW), PostgreSQL, MongoDB, Apache Spark
MLOps & Cloud: MLflow, Evidently, Prefect, Prometheus, Grafana, GitHub Actions CI/CD, GCP (Vertex AI, Cloud Run, Pub/Sub), Linux
Backend & Systems: Microservices, System Design, Auth & RBAC, REST, WebSockets, Node.js, Linux Kernel Modules

EDUCATION

B.Tech, Computer Engineering Jul 2022 – Jun 2026
Vishwakarma Institute of Technology, Pune CGPA 8.03

PUBLICATIONS & ACHIEVEMENTS

- **Papers:** Efficient Information Extraction from Large PDFs Using RAG and LLMs (Springer Nature) · Enhanced Railway Safety System (IEEE) · Alzheimer’s Diagnosis from Structural MRI (IJRASET) · Social Distancing ID Card (Scopus-indexed)
- **eYantra, IIT Bombay:** AIR 1 in Stage 1; top 10 worldwide in Stage 2
- **Certifications:** Fundamentals of Deep Learning (NVIDIA) · NLP & Deep Learning (DeepLearning.AI) · Google Cloud Computing Foundations