Raghav Kachroo

(858)-241-1760 | rkachroo@ucsd.edu | linkedin.com/raghavkachroo | github.com/mister-raggs

Software Engineer drawn to friction points where systems are slow, manual, or breaking under load - building high-throughput data systems and ML infrastructure that fixes them structurally.

Education

University of California, San Diego Sep 2024 – Mar 2026
Master of Science in Data Science (Artificial Intelligence & Machine Learning)

Indraprastha Institute of Information Technology, Delhi Sep 2022 – Sep 2023
Post Graduate Diploma in Data Science & Artificial Intelligence

University of Delhi Jul 2018 – Jun 2021
Bachelor of Management Studies

Experience

Amazon Jun 2025 – Sep 2025
Software Development Engineer Intern Bellevue, WA

- Built a distributed log indexing and query service over **42M+** log entries/hour using parallelized binary search, reducing incident triage latency from **15+ minutes to under 45 seconds**.
- Automated on-call SOPs using AWS Step Functions and Lambda, saving **12+ engineer-hours per week**.
- Integrated the log query service with internal diagnostic tooling via MCP, adding caching and query batching to maintain **sub-2s** response times under concurrent incident response load.

Hao AI Lab, UC San Diego Jan 2026 – Present
Student Researcher San Diego, CA

- Shipped **15.6%** wall-time reduction on Wan2.1-T2V-1.3B inference by identifying a deferred fp32 GPU to CPU tensor copy post-decode; quantized to uint8 on-device, gated fp32 on `return_frames=True`.
- Eliminated FA2/FA3 `flash_attn_func` graph break under `torch.compile` - **3.0%** compile median, **13.5%** warmup; extending `register_autograd` across varlen paths for training-under-compile.

Aark Global Apr 2023 – Sep 2024
Software Developer, AI/ML Delhi, India

- Built an async document ingestion pipeline processing **18,000+ pages/day** via Azure Queue Storage worker pool, maintaining throughput under variable load.
- Implemented read/write routing during datastore migration - Cosmos DB reads, MongoDB writes - maintaining **sub-100ms P95** latency throughout cutover.
- Built a full-text search pipeline ingesting scanned PDFs through OCR into indexed Elasticsearch documents, enabling **sub-180ms** query latency over previously unsearchable content.

Concentrix Jun 2022 – Mar 2023
Data Engineer Gurugram, India

- Replaced sequential scrapers with Airflow-orchestrated Kafka ingestion, increasing throughput by **60%** and cutting data freshness lag from **3 days to 6 hours**.

Projects & Research

Flare: ML-Powered Anomaly Detection Pipeline | *GitHub* Jan 2026 – Present

- Built an end-to-end ML anomaly detection pipeline applying Drain3 template mining, Isolation Forest scoring, and DBSCAN clustering to **11.2M log lines** across **575K blocks**, achieving **0.642 F1** at ∼**40ms** CPU-only latency.
- Built an LLM-as-judge eval harness grading incident summaries on relevance, specificity, and actionability, scoring **4.67/5** mean quality without requiring ground truth labels.

log-triage-v2: High-Throughput Log Indexing Engine | *GitHub* Jan 2026 – Apr 2026

- Go log indexing engine with time-sorted in-memory index, concurrent write ingestion, and binary search reads - sustains **42M rows/hour** with **sub-millisecond** nearest-timestamp lookup.

TRAI: AI Mobile App to Reduce the Gap Between Triage and Care | *IEEE* Jul 2025

- Deployed a clinical decision support service exposing ML inference via REST APIs, reducing inference time from **16s to 5s** through caching and request batching.

Technical Skills

Languages: Python, Go, SQL, Java
Backend & Infra: FastAPI, REST APIs, Docker, Kubernetes, CI/CD, GitHub Actions, Linux, MCP
Distributed & Data Systems: Kafka, Spark, Airflow, Elasticsearch, MongoDB, Cosmos DB, Azure Queue Storage, Redis
Observability & Cloud: Prometheus, OpenTelemetry, AWS (Lambda, Step Functions, S3), Azure
AI & ML: PyTorch, Scikit-Learn, Hugging Face Transformers, LangChain, MLflow, RAG Pipelines, Diffusers, LoRA
Concepts: Distributed Systems, Concurrency, MLOps, Inference Optimization, System Design, Model Monitoring