

RAGHAV KACHROO

(858)-241-1760 | rkachroo@ucsd.edu | linkedin.com/raghavkachroo | [github.com](https://github.com/rkachroo)

EDUCATION

University of California, San Diego	Sep 2024 – Mar 2026
Master of Science in Data Science (Artificial Intelligence & Machine Learning)	
• Relevant Coursework: Distributed ML Systems, LLM & AI Agents, Data Structures & Algorithms	
Indraprastha Institute of Information Technology, Delhi	Sep 2022 – Sep 2023
Post Graduate Diploma in Data Science & Artificial Intelligence	
University of Delhi	July 2018 – June 2021
Bachelor of Management Studies	

EXPERIENCE

Amazon	June 2025 – Sep 2025
Software Development Engineer Intern	Bellevue, WA
• Engineered an incident resolution toolkit to cut 12+ engineering hours per week by orchestrating AWS Step Functions and DynamoDB to automatically enrich incident tickets and execute routine SOPs.	
• Architected a search microservice to process 5GB+ of distributed logs in under 45 seconds by utilizing AWS Lambda to concurrently filter streams based on time and keyword inputs.	
• Integrated infrastructure log retrieval into chat workflows via a Model Context Protocol (MCP) server, enabling LLM-powered analysis of production logs through natural language.	
• Established reusable SDK and MCP components to standardize LLM agent development and evaluation workflows across 7 teams.	
Aark Global	Apr 2023 – Sep 2024
Software Developer, AI/ML	Delhi, India
• Orchestrated a high-throughput pipeline processing 3,000+ documents daily by utilizing Azure Queue Storage to distribute workloads across a pool of Virtual Machines.	
• Designed a hybrid data layer to deliver sub-100ms P95 latency across production requests by bridging MongoDB for rapid development with Cosmos DB for high performance serving.	
• Architected a full text search capability by designing a Tesseract OCR workflow to populate an Elasticsearch inverted index, delivering query results in under 180ms .	
• Integrated a conversational AI interface over indexed data to enable LLM-driven document analysis and question answering.	

Concentrix	Jun 2022 – Mar 2023
Data Engineer	Gurugram, India
• Built a Sentence-BERT and scikit-learn classification pipeline to automate sentiment analysis, reducing turnaround time from 3 days to 6 hours .	
• Scaled data ingestion to increase brand coverage by 60% by replacing legacy scraping with parallelized Airflow and Kafka pipelines capable of monitoring 8 major social and e-commerce platforms.	
• Implemented proactive error detection and logging for ingestion workflows, reducing resolution time by 67% .	

RESEARCH & PROJECTS

LLM Finetuning for Patient Routing + Clinical Decision Support https://ieeexplore.ieee.org/document/11252667
• Developed a patient triage service using BioBERT on AWS Bedrock + FastAPI, achieving 91.6% accuracy across 1,000+ assessments with EHR-integrated REST APIs.
• Optimized pipelines to cut inference time from 16s to 5s by using LangChain routing logic and response caching.

Autonomous GPU Rental Agent GitHub
• Built an autonomous LLM-driven system to plan and execute multi-step GPU rental workflows including vendor selection, budgeting, job execution, monitoring, retries, and payment handling.
• Architected a MongoDB-backed state machine enabling multi-hour workflows with structured planning, crash-safe recovery, and automatic vendor switching across serverless restarts.

TECHNICAL SKILLS

AI & ML: PyTorch, TensorFlow, Transformers, LangChain, Hugging Face, Scikit-learn, LlamaIndex, LangGraph

LLM Systems: Agent SDKs, Model Context Protocol (MCP), RAG, Vector Databases

Programming: Python, Java, JavaScript, REST APIs, System Design

Data Engineering: SQL, MongoDB, Neo4j, ETL Pipelines, Apache Kafka, Data Modeling, Monitoring

Cloud & Infra: AWS (Lambda, Bedrock, S3, EC2, CloudWatch), Azure, GCP, Docker, CI/CD, Git, Distributed Systems