

# RAGHAV KACHROO

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

## EDUCATION

<b>University of California, San Diego</b>	Sep. 2024 – Mar. 2026
Master of Science in Data Science (Artificial Intelligence & Machine Learning)	GPA: 3.63/4.0
• Relevant Coursework: Distributed ML Systems, LLM & AI Agents, Data Structures & Algorithms	
<b>Indraprastha Institute of Information Technology, Delhi</b>	Sep. 2022 – Sep. 2023
Post Graduate Diploma in Data Science & Artificial Intelligence	GPA: 4.0/4.0
<b>University of Delhi</b>	Jul. 2018 – Jun. 2021
Bachelor of Management Studies	GPA: 3.57/4.0

## EXPERIENCE

<b>Amazon</b>	Jun. 2025 – Sep. 2025
Software Development Engineer Intern	Bellevue, WA
• Engineered an incident resolution toolkit to cut <b>12+ engineering hours</b> per week by orchestrating AWS Step Functions and DynamoDB to automatically enrich incident tickets and execute routine SOPs.	
• Architected a search microservice to process <b>5GB+</b> of distributed logs in <b>under 45 seconds</b> by utilizing AWS Lambda to concurrently filter streams based on time and keyword inputs.	
• Integrated log retrieval tools into chat workflows via a <b>Model Context Protocol (MCP)</b> server, enabling developers to query infrastructure logs directly via natural language.	
• Established the blueprint for AI Agent development by establishing reusable SDK and MCP components for 7 teams.	
<b>UC San Diego</b>	Sep 2025 – Dec 2025
Teaching Assistant	San Diego, CA
• Designed course curriculum to bridge theory and practice by developing 4 coding assignments focused on RAG pipelines, Multi-Agent Orchestration, and LLM Finetuning using LangChain and PyTorch.	
<b>Aark Global</b>	Apr. 2023 – Sep. 2024
Software Developer, AI/ML	Delhi, India
• Orchestrated a high-throughput pipeline processing <b>3,000+ documents daily</b> by utilizing Azure Queue Storage to distribute workloads across a pool of Virtual Machines.	
• Designed a hybrid data layer to deliver <b>sub-100ms P95 latency</b> 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 <b>under 180ms</b> .	
• Integrated a conversational AI interface over the indexed data, enabling users to perform high level document analysis and interact with complex datasets through natural language.	

<b>Concentrix</b>	Jun. 2022 – Mar. 2023
Data Engineer	Gurugram, India
• Automated sentiment analysis workflows to reduce analysis turnaround time from <b>3 days to 6 hours</b> by replacing manual labeling with a Sentence-BERT and scikit-learn classification pipeline.	
• Scaled data ingestion to increase <b>brand coverage by 60%</b> 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 <b>resolution time by 67%</b> and improving reporting reliability.	

## RESEARCH & PROJECTS

<b>LLM Finetuning for Patient Routing + Clinical Decision Support</b>   <a href="https://doi.org/10.1016/S0016-5085(25)04641-4">doi.org/10.1016/S0016-5085(25)04641-4</a>
• Developed a patient triage service using BioBERT on AWS Bedrock + FastAPI, achieving <b>91.6%</b> accuracy across <b>1,000+</b> assessments with EHR-integrated REST APIs.
• Optimized pipelines to cut inference time from <b>16s to 5s</b> by using LangChain routing logic and response caching.

## TECHNICAL SKILLS

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

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

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

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

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