RAHUL SHETTY

United States | rahulshetty61199@gmail.com | https://github.com/RaSh6119 | +1 (720) 803-9791

EDUCATION

University of Colorado, Boulder - Master of Science, Computer Science | GPA: 3.8/4.0 Aug 2024 – May 2026
- **Coursework:** Machine Learning, Natural Language Processing, Datacenter Scale Computing, Distributed Systems

Sardar Patel Institute of Technology - Bachelor of Technology, ECE | GPA: 9.1/10 Aug 2018 – May 2022
- **Coursework:** Data Structures and Algorithms, Programming (Python, Java), Machine Learning, Signals

SKILLS

Languages: Python, Java, SQL, MATLAB, BASH scripting, JavaScript, HTML5, CSS3 | **ML & AI:** TensorFlow, PyTorch, Pandas, Matplotlib, GenAI, CNN, Transformers, RAG, Knowledge Graphs | **Frameworks & Libraries:** ReactJS, ROS2, NumPy, LangChain, Hadoop, Spark, Streamlit | **Technologies:** Ansible, Docker, GitHub, Linux, ChromaDB, Pinecone, Kubernetes, Firebase

EXPERIENCE

Graduate Researcher, Human Interaction and Robotics Group – CU Boulder Jan 2025 – Present
- Developed **DEFT, a diffusion-based generative AI framework** that learns feasible robot trajectories under **single and multi-joint failures**, improving constraint satisfaction from **37%** → **75%** and transport task success from **42%** → **99%**.
- Designed **embodiment and task-conditioned diffusion models (FiLM-conditioning)** enabling zero-shot adaptation to unseen failure conditions without retraining, achieving **74%** success on **out-of-distribution failures**.
- Built an **ML pipeline** for data generation, model training, and constraint enforcement, and validated on **4.7M+ trajectories across 2,400+ failure scenarios** plus a real-world multi-primitive drawer task, demonstrating robust fail-active autonomy.
- Manuscript submitted to **IEEE International Conference on Robotics and Automation (ICRA) - Under Review**.

Software Development Engineer, Nomura July 2022 – July 2024
- Enhanced the Front Office Supervision (FOS) portal by creating an **organizational hierarchy tree** from data fetched from the **Oracle** database and refactored parts of the **SpringBoot** code.
- Collaborated with a team to migrate **700+ data marts (13TB) from Sybase to Snowflake on AWS**, saving over **$800K in infrastructure costs** while modernizing **three large-scale applications**.
- Led the development of a **Snowpark framework** using **Python, Shell scripting and SnowSQL** to streamline **Informatica** workflows and cut data processing time by **75%**.

Software Engineering Intern, Nomura Jan 2022 – June 2022
- **Automated** retrieval of **Git** details for production releases through **JIRA API integration**, streamlining release preparation and **eliminating human errors**. Developed a scalable **Python** backend with a **Django** web application frontend, supporting **10+ teams** across the organization.
- **Led the migration** of an existing **JSP**-based monitoring portal to a React.js frontend and a refactored **SpringBoot** backend, resulting in a **100%** code overhaul.

PROJECTS

Agentic Hybrid RAG Engine | | Python, Neo4j, Qdrant, LangChain, Docker, Streamlit Dec 2025 - Jan 2026
- Architected a **Neuro-Symbolic RAG system** integrating **Knowledge Graphs (Neo4j)** and **Vector Search (Qdrant)**, solving "**multi-hop**" **reasoning failures** where standard **LLM retrieval scored 0/10**.
- Engineered a **Semantic Router** to **dynamically dispatch queries**, achieving a **106% accuracy improvement (8.25/10 vs 4.0/10)** over naive **vector baselines** and outperforming state-of-the-art **HyDE methods**.
- Reduced **hallucination rates to 0%** on complex entity-relationship tasks by grounding responses in graph structures, validated through a custom **LLM-as-a-Judge benchmarking pipeline**.

Photo Memory Finder | | Python, GCP (Vertex AI, Cloud Run), React, Pinecone Nov 2025 – Dec 2025
- Architected a serverless, event-driven cloud application on **Google Cloud Platform** capable of processing concurrent photo uploads with **<2s latency**, achieving semantic understanding beyond simple metadata.
- Built a **multimodal** ingestion pipeline using **Cloud Pub/Sub** and **Cloud Run** workers to asynchronously process uploads, reducing metadata extraction time by **40%** via parallelized **hybrid embedding generation (image + text)** with **Vertex AI**.
- Implemented an intelligent LLM Reranking system using **Gemini 2.5 Flash**, filtering and re-ordering vector search results based on query context, improving **retrieval precision by 35%** for complex queries.

EXTRACURRICULARS

- **Runner-up, HackCU 11 (AMD AI Track)** – Secured 2nd place among **50+ teams** with an AI-powered application on AMD Ryzen AI hardware.