RISHABH JAIN

(646)484-0525 rj2790@columbia.edu r-rishabh-j.github.io linkedin.com/in/rishabhj11 github.com/r-rishabh-j

EDUCATION

Columbia University, New York Aug 2025 - Dec 2026
Master of Science (MS) Computer Science (Machine Learning) *New York, NY*
- Machine Learning, High Performance ML, Natural Language Processing, Computational Aspects of Robotics

Indian Institute of Technology (IIT) Ropar Jul 2019 - May 2023
B.Tech (Honors) Computer Science and Engineering, Concentration in AI | CGPA 8.7/10, AI 8.93/10 *Ropar, IND*

PROFESSIONAL EXPERIENCE

Software Engineer, Arista Networks Jul 2023 - Jun 2025
C, C++, Python, Docker, Software Defined Networking *Bengaluru, India*
- Engineered low-latency data and control plane components in C++ for the EOS software forwarding engine, enhancing throughput and processing efficiency for high-volume packet processing and telemetry across 16 core repositories
- Architected scalable state management agents and synchronization modules in C++ and Python to support up to 90 million entries in the EOS concurrent packet flow hash table
- Led the creation of an automated build orchestration tool that resolves complex dependency graphs for upstream AlmaLinux packages, streamlining workflows for 15+ teams during a company-wide transition from P4 to git

Edison AI Intern, General Electric Healthcare May 2022 - Jul 2022
Python, PyTorch, FastAPI, PostgreSQL, Docker *Bengaluru, India*
- Created a real-time patient monitoring pipeline in Edison Digital Health Platform using the YOLOv5 model
- Developed a lightweight model through ablation, and fine-tuned it on open-source and over 30,000 self-annotated images. Deployed it through containerized APIs using FastAPI backend and PostgreSQL database

RESEARCH WORK AND PROJECTS

Aligning LLMs for Speculative Decoding via Task-Adaptive Knowledge Distillation Oct 2025 - Dec 2025
PyTorch, LLMs, Speculative Decoding, Knowledge Distillation
- Implemented a custom Speculative Decoding framework supporting dynamic batching and non-uniform draft lengths
- Performed white-box, token-level On-Policy Knowledge Distillation to align low-cost draft models from Qwen3, SmolLM families with larger target models, effectively mitigating exposure bias to accelerate speculative generation
- Benchmarked token and sequence level acceptance rates over various divergence objectives (Forward/Reverse KL, JSD), achieving a 5% increase in token acceptance rate after just 1 epoch of distillation on GSM8k and 4% on CNN-DM

NFR Benchmarking for AI Agents in IBM ITBench Oct 2025 - ongoing
ITBench, AI Agents, CrewAI, Langfuse | GitHub: ITBench-NFR *IBM Research, Columbia University*
- Co-developed a non-functional requirements evaluation framework extending ITBench, defining a comprehensive two-level taxonomy for agent-specific requirements (cost efficiency, reliability, observability) and instrumenting SRE, CISO and Mini-SWE agents with Langfuse, OpenInference, and vLLM for granular telemetry
- Conducted comparative evaluations across ReAct and Plan&Execute architectures on 15 SRE incidents and 3 CISO scenarios using Gemini and Qwen LLMs, revealing Plan&Execute agents achieved up to 15x higher Prompt-to-Completion Ratio and significantly lower latency than ReAct

Viewpoint-Invariant Robot Manipulation via 3D Geometric Priors Oct 2025 - Dec 2025
PyTorch, Mujoco, Gymnasium *Columbia University*
- Developed a hybrid transformer-based model that fuses PointNet-encoded 3D priors with egocentric 2D features to mitigate covariate shift from view-point perturbations in imitation learning
- Performed extensive ablations to demonstrate that hybrid egocentric cues are crucial for contact-rich tasks
- Demonstrated zero-shot generalization to novel viewpoints, increasing success rates from 0% to ∼70% by effectively separating global geometric structure from local semantic appearance

Video Transformer Based Multi-view Body Language and Behaviour Recognition May 2023 - Oct 2023
Python, PyTorch, Deep Learning, Computer Vision *Monash University*
- Built a multi-view feature-fusion pipeline with a finetuned VideoSwin transformer backbone for multi-label classification
- Placed 2nd in the ACM MultiMedia 2023 Bodily Behaviour Recognition Grand Challenge
- Published work at ACM MultiMedia 2023 and IEEE Transactions on Affective Computing

TECHNICAL SKILLS

Languages: C, C++, Python, Java
Math & AI: NumPy, OpenCV, PyTorch, Gym, MuJoCo
Database & Backend: PostgreSQL, PostGIS, FastAPI
Tools: Git, Perforce, Bash, Docker, HuggingFace, vLLM