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) | GPA: 3.92/4 *New York, NY*

- ML Decision Theory, High Performance ML and Inference, Natural Language Processing, Robot Manipulation

Indian Institute of Technology Ropar Jul 2019 - May 2023
B.Tech (Honors) Computer Science and Engineering, Concentration in Artificial Intelligence *Ropar, IND*

- Algorithms, Databases, Operating Systems, Data Science, Neural Networks, Advanced Computer Vision

TECHNICAL SKILLS

Languages: C, C++, Python, Java, RISC-V, Bash
Backend: PostgreSQL, PostGIS, FastAPI, Google Cloud
Math & AI: NumPy, Pandas, CUDA, PyTorch, HuggingFace, MuJoCo
Tools: Git, Linux, Docker, vLLM, Langfuse, OpenAI SDK

PROJECTS AND RESEARCH

Accelerating Speculative Decoding for LLMs via On-Policy Knowledge Distillation Oct 2025 - Dec 2025
PyTorch, Huggingface TRL | GitHub: r-rishabh-j/batched_specdec, r-rishabh-j/distillSpec

- Implemented a speculative decoding engine with batching, non-uniform acceptance length and kv-cache pruning
- Distilled Qwen3-0.6B and SmolLM-360M drafters from Qwen3-4B and SmolLM-1.7B via logit On-Policy Distillation
- Benchmarked token and sequence level acceptance rates over FKL, RKL and JS divergence objectives, achieving 5% increase in token acceptance rate on GSM8k and 4% on CNN-DM

NFR Benchmarking in IBM ITBench for IT Automation AI Agents Oct 2025 - present
CrewAI, Langfuse, vLLM | GitHub: ITBench-NFR *IBM Research, Columbia University*

- Co-developing a non-functional requirements (NFR) evaluation framework extending ITBench - defining a two-level taxonomy for agent-specific requirements and instrumenting SRE, CISO and Mini-SWE agents with Langfuse and vLLM
- Compared ReAct and Plan&Execute architectures on ITBench scenarios using Gemini-2.5-Pro and Qwen3-14B LLMs

Viewpoint-Invariant Robot Manipulation via 3D Geometric Priors Nov 2025 - Dec 2025
PyTorch, MuJoCo, Gymnasium, Robotics | GitHub: r-rishabh-j/3DEgoACT

- Enhanced ACT to fuse 3D point-cloud with egocentric 2D views to mitigate inference-time view-point perturbations in low-cost robot imitation learning based policies. Achieved zero-shot generalization ($\sim$70%) where baseline collapsed

Video Transformer Based Multi-view Body Language and Behaviour Recognition May 2023 - Oct 2023
PyTorch, Computer Vision | GitHub: MAGIC-TBR *Monash University, IIT Ropar*

- Built a multi-view feature-fusion VideoSwin transformer based pipeline for multi-label classification of body behavior
- Placed 2nd in the ACM MM 2023 Grand Challenge, published papers at ACM MM 2023 and IEEE TAFFC

COMPETITIONS

- **Jan 2026**: **Won Qualcomm Snapdragon Multiverse Hackathon** by building a multi-device multi-modal AI coding assistant on Snapdragon devices supporting user input beyond text (voice, stylus annotations)
- **Jul 2023**: Runner up in **ACM MultiMedia Grand Challenge** for building a multi-view classification model for videos

WORK EXPERIENCE

Software Engineer, Arista Networks Jul 2023 - Jun 2025
C, C++, Python, Linux, Git, Perforce *Bengaluru, India*

- Worked on low-level BESS & DPDK C++ modules in EOS software forwarding engine for optimizing throughput and memory efficiency. Led creation of stateful bi-directional flow modules in CloudVision IPFIX across ICMP, TCP, GRE stacks
- Scaled flow monitoring backend to support a 9× capacity increase in the EOS concurrent flow hash table. Re-wrote flow scale test suites with multiprocessing in Python, achieving a 6× gain in evaluation throughput with 90 million flows
- Created RPM build tools to resolve upstream AlmaLinux dependency graphs with Arista patches during a company-wide shift from Perforce mono-repo to Git multi-repo. Streamlined development workflows for 15+ teams

Edison AI Intern, General Electric Healthcare Jun 2022 - Jul 2022
PyTorch, FastAPI, PostgreSQL, Docker, Computer Vision *Bengaluru, India*

- Developed a vision based system for real-time, face-indexed spatio-temporal tracking of admitted patients in hospitals
- Co-annotated an internal dataset, ablated YOLOv5 to develop a low-cost model, and fine-tuned it with just 5GB VRAM