

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

- Spring'26: ML Decision Theory, Robot Manipulation, Quantum Computing
- Fall'25: ML Theory, High Performance ML, NLP, Computational Aspects of Robotics

Indian Institute of Technology Ropar

July 2019 - May 2023

B.Tech (Honors) Computer Science and Engineering, Concentration in Artificial Intelligence

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

WORK EXPERIENCE

Software Engineer, Arista Networks

Jul 2023 - Jun 2025

C, C++, Python, Docker, Software Defined Networking

Bengaluru, India

- Worked on low-level BESS & DPDK C++ modules in EOS software forwarding engine for scaling packet processing throughput and memory access efficiency
- Led creation of stateful bi-directional flow modules in CloudVision IPFIX across ICMP, TCP, GRE stacks
- Contributed in building support for a 9× capacity increase in the EOS concurrent flow hash table. Re-wrote flow table scale test suite with multiprocessing in Python, achieving a 6× gain in evaluation throughput
- Designed configuration CLIs and SysDB agents to support for MSS firewall in EOS network switches
- Created internal RPM build tools to resolve upstream AlmaLinux dependencies with Arista patches during a company-wide shift from Perforce mono-repo to Git multi-repo. Streamlined development workflows for 15+ teams with rapid adoption and contributions within 1 week of release

Edison AI Intern, General Electric Healthcare

May 2022 - Jul 2022

Python, PyTorch, FastAPI, PostgreSQL, Docker

Bengaluru, India

- Developed a computer vision based system for real-time, face-indexed spatio-temporal tracking of admitted patients in hospitals. Deployed it through containerized endpoints in the Edison Health platform
- Co-curated an internal dataset of 30,000 images, ablated YOLOv5 to develop a lightweight model, and fine-tuned it under resource-constraints of only 5GB of VRAM, taking over 280 GPU hours

PROJECTS AND RESEARCH

Accelerating Speculative Decoding via On-Policy Knowledge Distillation

Oct 2025 - Dec 2025

PyTorch, Huggingface TRL | GitHub: [r-rishabh-j/distillSpec](https://github.com/r-rishabh-j/distillSpec), [r-rishabh-j/batched_specdec](https://github.com/r-rishabh-j/batched_specdec)

Columbia University

- Implemented a speculative decoding engine with prompt batching, non-uniform acceptance length, batched verification, kv-cache pruning
- Distilled Qwen3-0.6B and SmoLM-360M drafters from Qwen3-4B and SmoLM-1.7B respectively via sequence level white-box On-Policy Knowledge Distillation to align models for accelerating speculation
- Benchmarked token and sequence level acceptance rates over Forward KL, Reverse KL and JS divergence objectives, achieving 5% increase in token acceptance rate after 1 epoch on GSM8k and 4% on CNNDM

NFR Benchmarking in IBM ITBench for IT Automation Agents

Oct 2025 - ongoing

PyTorch, ITBench, CrewAI, Langfuse, LLMs | GitHub: [ITBench-NFR](https://github.com/ITBench-NFR)

IBM Research

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

Viewpoint-Invariant Robot Manipulation via 3D Geometric Priors Python, PyTorch, Mujoco, Gymnasium GitHub: r-rishabh-j/3DEgoACT	Oct 2025 - Dec 2025 Columbia University
<ul style="list-style-type: none"> Modified ACT to take PointNet-encoded 3D point-cloud as input tokens along with egocentric 2D features to mitigate inference-time covariate shift from view-point perturbations in imitation learning policies Performed ablations to demonstrate that egocentric cues are crucial alongside allocentric 3D features for contact-rich tasks under a fixed training budget Demonstrated zero-shot generalization to perturbed viewpoints, achieving ~70% success rate in scenarios where standard ACT failed by effectively decoupling global structure from view-dependent appearance 	
Video Transformer Based Multi-view Body Behaviour Recognition Python, PyTorch, Deep Learning, Computer Vision	May 2023 - Oct 2023 Monash University & IIT Ropar
<ul style="list-style-type: none"> Built a multi-view feature-fusion pipeline with a finetuned VideoSwin transformer backbone for multi-label classification GitHub: MAGIC-TBR Published MAGIC-TBR: Multi-view Attention Fusion for Transformer based Bodily Behavior Recognition in Group Settings at ACM MultiMedia, 2023 Published Multi-view Attention Fusion for Explainable Body Language Behavior Recognition at IEEE TAFFC Placed 2nd in the ACM MultiMedia 2023 Bodily Behaviour Recognition Grand Challenge certificate 	
Spatio-Temporal Hotspot Detection in Microsoft Azure Java, Python, PostgreSQL, PostGIS document	Aug 2022 - Nov 2023 Microsoft & IIT Ropar
<ul style="list-style-type: none"> Formulated a statistical framework to identify spatio-temporal hotspots in Microsoft Azure from network autonomous system data from 10+ Indian cities stored in a spatial PostGIS database Contributed to implementation of algorithms and database CRUD, creation of synthetic data, and testing on Microsoft's proprietary dataset Publication: Periodic Spatio-Temporal Colored Hotspot Detection in Azure Traffic Data; ACIIDS 2025 	
PUBLICATIONS	
1. Multi-view Attention Fusion for Explainable Body Language Behavior Recognition . <i>IEEE TAFFC</i> , 2025	
2. Periodic Spatio-Temporal Colored Hotspot Detection in Azure Traffic Data . <i>ACIIDS</i> , 2025	
3. MAGIC-TBR: Transformer-based Bodily Behavior Recognition in Group Settings . <i>ACM Multimedia</i> , 2023	
TECHNICAL SKILLS	
Languages: C, C++, Python, Java, RISC-V, Bash	
Tools: Linux, Git, Perforce, Docker, Google Cloud, Gemini & OpenAI API, Langfuse, vLLM, MuJoCo	
Libraries: Numpy, Pandas, FastAPI, OpenCV, CUDA, PyTorch, Gymnasium, Huggingface	
OTHER PROJECTS	
RFDN Variants: Efficient Image Super-Resolution NTIRE CVPR Challenge Python, PyTorch, Computer Vision document	Feb 2023-May 2023 LASII Lab, IIT Ropar
<ul style="list-style-type: none"> Developed efficient image super-resolution model variants of the CNN based RFDN baseline Studied trade-offs between accuracy and runtime among variants and achieved a superior PSNR on the DIV2K dataset along with a reduced model inference time 	
Client Selection in Deep Federated Recommender Systems Python, PyTorch, Machine Learning document	Oct 2022-March 2023 Game Theory Lab, IIT Ropar
<ul style="list-style-type: none"> Developed client subset selection strategies to optimize training costs in federated recommender systems Evaluated strategies over collaborative filtering based deep recommender systems on MovieLens datasets 	