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

- Data Science, Artificial Neural Networks, Artificial Intelligence, 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 C++ programming in BESS and DPDK EOS software forwarding engine modules
- Led creation of stateful flow sync modules in C++ across ICMP, TCP, GRE and CloudVision IPFIX 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 Arista MSS firewall in EOS network switches
- Created an internal RPM build tool 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 within 1 week of release

Edison AI Intern, General Electric Healthcare **May 2022 - Jul 2022**

Python, PyTorch, FastAPI, PostgreSQL, Docker *Bengaluru, India*

- Co-developed an end-to-end spatio-temporal patient tracking pipeline using YOLOv5 and PostgreSQL, allowing for the continuous storage and retrieval of patient location data. Deployed it through containerized FastAPI endpoints
- Curated an internal dataset of 30,000 images, ablated YOLOv5 to develop a lightweight model, and fine-tuned it under resource-constraints (5GB VRAM)

PROJECTS AND RESEARCH

Accelerating Speculative Decoding via On-Policy Knowledge Distillation **Oct 2025 - Dec 2025**

PyTorch, Huggingface TRL | GitHub: r-rishabh-j/distillSpec, r-rishabh-j/batched_specdec *Columbia University*

- Implemented a speculative decoding engine with prompt batching, non-uniform acceptance length, batched verification, kv-cache and pruning
- Distilled Qwen3-0.6B and SmolLM-360M drafters from Qwen3-4B and SmolLM-1.7B respecitvely 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 *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 **Oct 2025 - Dec 2025**
Python, PyTorch, Mujoco, Gymnasium | GitHub: *r-rishabh-j/3DEgoACT* *Columbia University*

- 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 $\sim$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 **May 2023 - Oct 2023**
Python, PyTorch, Deep Learning, Computer Vision *Monash University & IIT Ropar*

- 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 **Aug 2022 - Nov 2023**
Java, Python, PostgreSQL, PostGIS | *document* *Microsoft & IIT Ropar*

- 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

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 **Feb 2023-May 2023**
Supervised by Dr. Abhinav Dhall | *document* *LASII Lab, IIT Ropar*

- 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 **Oct 2022-March 2023**
Supervised by Dr. Shweta Jain | *document* *Game Theory Lab, IIT Ropar*

- 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

Dynamic Planning in Dyna-Q for Faster Training **Sept 2022 - Nov 2022**
Supervised by Dr. Shashi Shekhar Jha | *document* *IIT Ropar*

- Explored dynamic planning schedules in DYNA-Q& Deep DYNA-Q reinforcement learning algorithms.
- Evaluated trade-offs in performance and training costs of various schedules on OpenAI Gym environments.