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 with Concentration in AI *Ropar, IND*
- AI Concentration in Deep Learning, Computer Vision, Reinforcement Learning, Federated Learning

TECHNICAL SKILLS

Languages: C, C++, Python, Java **Math&AI**: Numpy, OpenCV, PyTorch, Gymnasium, Mujoco
Tools: Git, Perforce, Linux, Docker, Android SDK **Database&Backend**: PostgreSQL, PostGIS, Flask, FastAPI

RESEARCH WORK AND PROJECTS

Aligning LLMs for Speculative Decoding via Task-Adaptive Knowledge Distillation Nov 2025 - Dec 2025
Python, 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 in IBM ITBench for IT Agents Oct 2025 - ongoing
PyTorch, ITBench, CrewAI, Langfuse *IBM Research*
- 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 complex 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 | GitHub: ITBench-NFR

Video Transformer Based Multi-view Body Language and Behaviour Recognition May 2023 - Oct 2023
Python, PyTorch, Deep Learning, Computer Vision
- Publication: MAGIC-TBR: Multi-view Attention Fusion for Transformer based Bodily Behavior Recognition in Group Settings; ACM MultiMedia, 2023
- Publication: Multi-view Attention Fusion for Explainable Body Language Behavior Recognition; IEEE TAFFC
- 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 | certificate

Spatio-Temporal Hotspot Detection in Microsoft Azure | BTech Capstone Aug 2022 - Nov 2023
Java, Python, PostgreSQL, PostGIS *Microsoft*
- 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

PROFESSIONAL EXPERIENCE

Software Engineer, Arista Networks Jul 2023 - Jun 2025
C, C++, Python, Docker, Software Defined Networking *Bengaluru, India*
- Contributed to early stage feature development in control and data plane of the Multi-Domain Segmentation Service. Designed essential TACC constructs and Sysdb agents in 4 new repositories by analyzing future feature requirements
- Enhanced EOS Software Forwarding Engine features, throughput and performance across 16 core repositories
- Joined the EngProd team in January 2025. Proposed and led development of a new auto-build tool to streamline workflows of 15+ teams at Arista during a company-wide version control transition. Tool bundles external AlmaLinux package dependencies in Arista EOS by resolving dependency graphs across all of Arista RPMs

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