Devansh Bhardwaj

github.com/fireshadow05 ✉ bhardwajdevansh398@gmail.com

Education

Indian Institute of technology Roorkee Nov 2021 - Present
Bachelor of Technology in Electronics and Communication Engineering *Current CGPA: 8.20/10.0*

Skills and Interests

Languages and Tools: C/C++, Java, Python, Ros, LaTeX, Git/GitHub, Unix Shell
Libraries and Frameworks: Pytorch, Tensorflow, JAX, pandas, NumPy, Matplotlib, scikitlearn, RestAPI
Interests: Generative AI, Reinforcement learning, Adversarial Machine Learning, Multimodal Learning, Natural Language Processing,

Publications, Preprints and Workshops

Agent Context Protocols Enhance Collective Inference
Devansh Bhardwaj*, *Arjun Beniwal*, Shreyas Chaudhari, Ashwin Kalyan, Tanmay Rajpurohit, Karthik R Narasimhan, Ameet Deshpande, Vishvak Murahari* Arxiv Project Page
Preprint

One Noise to Fool Them All: Universal Adversarial Defenses Against Image Editing
Devansh Bhardwaj*, *Parth Badgujar*, Shorya Singhal** Link
CVPR 2025 Advml Workshop

Rethinking Randomized Smoothing from the Perspective of Scalability
Anupriya Kumari,* ***Devansh Bhardwaj****, *Sukrit Jindal*, Sarthak Gupta* Arxiv
NeurIPS AdvML-Frontiers'24 Workshop

Invisible Traces: Using Hybrid Fingerprinting to identify underlying LLMs in GenAI Apps
Devansh Bhardwaj, *Naman Mishra* Arxiv
Preprint

Video Generation with Learned Action Prior
Meenakshi Sarkar, ***Devansh Bhardwaj***, *Debasish Ghose* Arxiv Code
Preprint

Accelerated Smoothing: A Scalable Approach to Randomized Smoothing
Devansh Bhardwaj, *Kshitiz Kaushik, Sarthak Gupta* Arxiv Code
Preprint

Experience

Princeton University | *Undergraduate Researcher* June 2024 - Present

- At Prof. Karthik Narsimhan's Lab I worked on Co-developing Agent Context Protocols (ACPs) for generalized and robust multi-agent communication and coordination.
- Designed Execution Blueprints (DAGs) enabling fault-tolerant, long-horizon workflows.
- Achieved SOTA results on AssistantBench (28.3%) and led evaluations across complex tasks including multimodal report generation and dashboard creation.

Repello.ai | *ML Security Researcher* March 2024 - Present

- Worked on developing a completely new paradigm and methodology for LLM Fingerprinting by only observing their outputs.
- Discovered innovative jailbreaking strategies for LLM applications and authored **blogs** recognized as key resources, including one featured in the **OWASP Top 10 for LLM Applications (2025).**
- Conducted red-teaming assessments on 10+ AI applications, gaining hands-on experience in identifying and mitigating security vulnerabilities.
- Worked on developing set of Guardrails to mitigate issues in LLM Applications such as Prompt Injection, Output Moderation, System Prompt Leak Detection, etc.

Indian Institute of Science, Bangalore | *Undergraduate Researcher* | Hybrid Oct 2023 – July 2024

- At Professor Debasish Ghose's Lab, I Worked on setting up an environment in **NVIDIA's ISAAC Sim**, a realistic robotics simulator, followed by the validation of an algorithm that utilizes Kalman filters and Bayesian belief spaces for human-intent detection.
- Collaborated on the development of the RoAM dataset, featuring 50 long video-action sequences, and benchmarked existing SOTA video prediction models while introducing 3 new baselines to improve dataset usability.

Projects

Behaviour and Content Simulation | *Inter-IIT Tech Meet 12.0* | *Code* Nov 2023 - Dec 2023

- Led development of a multi-modal LLM pipeline, involving two tasks, first to simulate behavior (likes) from the content of a tweet, second to simulate content (tweet text) from the tweet metadata.
- Transformed visual data from over 300000 images and videos into structured textual information utilizing BLIP-2 which lead to increase in the quality of tweet text generated.
- Conducted detailed experimentation with emerging methodologies such as in-context learning; insights from these experiments led to improved ROUGE-2 scores by 58% over baseline.

AIKavach | *DSG. IITR* | *Code* March 2023 – May 2023

- Developed a Flask-based web app to certify robustness of deep learning models against adversarial attacks.
- Enhanced user-provided model weights by integrating a denoiser for enhanced robustness.
- Integrated Input specific sampling to Double Sampling Randomized smoothing for faster robust radius certification.

Achievements and Extra Co-curricular

Silver Medal, Inter-IIT Tech Meet 12.0
Bronze Medal, Inter-IIT Tech Meet 11.0
All India Rank 1410 in Jee Advanced 2021

Data Science Group | *Secretary* June 2022 – Present

- Led a student group promoting Machine Learning through research paper discussions, and events with industry leaders like World Wide Technologies and Pathway. Conducted a workshop which received participation of over 800+ students.