Nirav **Diwan**

Ph.D. student in C.S., University of Illinois at Urbana-Champaign

https://nirav0999.github.io/ ndiwan2@illinois.edu github.com/nirav0999 Google Scholar Champaign, USA

Education

Aug 2024 - | University of Illinois Urbana-Champaign (UIUC) | Champaign, US
Ph.D. in Computer Science
Area: Security & Privacy, Machine Learning

Aug 2022 May 2024 | University of Illinois Urbana-Champaign (UIUC) | Champaign, US
Master of Science (M.S.) in Computer Science (CS) (thesis-track)
Specialization: Machine Learning

Aug 2017 Jun 2021 | Indraprastha Institute of Information Technology, Delhi (IIITD) | Delhi, India
Bachelor in Technology (B.Tech) in Computer Science & Engineering (CSE)
Dean's Award for Academic Excellence

Selected Publications

S=In Submission, C=Conference, W=Workshop, P=Poster/Demo, J=Journal, *Equal Contribution

[S.2] **PurpCode: Reasoning for Safer Code Generation** [🔗] [GitHub]
Jiawei Liu*, Nirav Diwan*, Zhe Wang*, Haoyu Zhai, Xiaona Zhou, Kiet A. Nguyen, Tianjiao Yu, Muntasir Wahed, Yinlin Deng, Hadjer Benkraouda, Yuxiang Wei, Lingming Zhang, Ismini Lourentzou, Gang Wang
39^{th} *The Annual Conference on Neural Information Processing Systems* **[NeurIPS 2025]**
Won the 1^{st} place in the Amazon Nova AI Challenge

[S.1] **MOCHA: Are Code Language Models Robust Against Multi-Turn Malicious Coding Prompts?** [🔗]
Muntasir Wahed, Xiaona Zhou, Kiet A. Nguyen, Tianjiao Yu, Nirav Diwan, Gang Wang, Dilek Hakkani-Tür, Ismini Lourentzou
30^{th} *Empirical Methods in Natural Language Processing (Findings), 2025* **[EMNLP 2025]**

[W.1] **Clear Preferences Leave Traces: Reference Model-Guided Sampling for Preference Learning** [🔗]
Nirav Diwan, Tolga Ergen, Dongsub Shim, Honglak Lee
1^{st} *Workshop on Preparing Good Data for Generative AI Challenges and Approaches, 2025* **[Good Data Workshop @ AAAI 2025]**

[C.4] **You Can't Judge a Binary by Its Header: Data-Code Separation for Non-Standard ARM Binaries using Pseudo Labels** [🔗]
Hadjer Benkraouda, Nirav Diwan, Gang Wang
46^{th} *IEEE Symposium on Security and Privacy, 2025* **[IEEE S&P 2025]**

[C.3] **It Doesn't Look Like Anything to Me: Using Diffusion Models to Subvert Visual Phishing Detectors** [🔗] [GitHub]
Qingying Hao, Nirav Diwan, Ying Yuan, Mauro Conti, Giovanni Apruzzese, Gang Wang
33^{rd} *USENIX Security Symposium, 2024* **[USENIX Security 2024]**

[C.2] **Weakening the Inner Strength: Spotting Core Collusive Users in YouTube Blackmarket Networks** [🔗] [GitHub]
Hridoy Sankar Dutta*, Nirav Diwan*, Tanmoy Chakraborty
16^{th} *International AAAI Conference on Web and Social Media, 2022* **[ICWSM 2022]**

[C.1] **Fingerprinting Fine-tuned Language Models in the Wild** [🔗] [GitHub]
Nirav Diwan, Tanmoy Chakraborty, Zubair Shafiq
59^{th} *Annual Meeting of the Association for Computational Linguistics (Findings), 2021* **[ACL 2021]**
Dean's Thesis Appreciation Award

[J.1] **RecipeDB: A Resource for Exploring Recipes** [🔗] [GitHub]
Devansh Batra*, Nirav Diwan*, Utkarsh Upadhyay*, Jushaan Singh Kalra*, Tript Sharma*, Aman Kumar Sharma*, Dheeraj Khanna*, Jaspreet Singh Marwah*, Srilakshmi Kalathil*, Navjot Singh*, Rudraksh Tuwani*, Ganesh Bagler*
Database: The Journal of Biological Databases and Curation, Oxford University Press, 2020 **[Database 2020]**
Press: Times of India, The National, Nature India

Awards

Amazon Nova AI Challenge (Winner): Won 1^{st} place as team co-lead in the Amazon Nova AI Challenge winning a cash prize of $250K.

Amazon Nova AI Grant: Awarded grants worth $**250, 000** + $**1.5 Million AWS Credits**.

Catalyzing Advocacy in Science and Engineering (CASE) Workshop 2024: Only student selected to represent the CS department.

Dean's Thesis Appreciation Award, IIITD: For *outstanding thesis research* from Dean of Academic Affairs.

Dean's Award for Academic Excellence, IIITD: For *exceptional academic performance* for the academic years 2019-20.

IIT-JEE 2017: Ranked 1884/1.5 million (top 0.15%) in IIT-JEE Mains and 4186/200,000 (top 2%) in IIT-JEE Advanced.

Recent Industrial Experience

1. LG AI Research | **Bilingual LLM Team** [🌐] — Ann Arbor, US

Research Intern | *Area: Natural Language Processing (NLP), Safety and Alignment* — 2024

- Improved alignment of LLMs for single-turn and multi-turn settings. Results published in the **AAAI Good Data Workshop**.
- Developed a sampling method that outperformed alignment methods (DPO, ORPO, SimPO) for Coding, Math, and Reasoning.

2. Ema Inc. | **Generative AI Team** [🌐] — Remote, US

Applied Science Intern | *Area: Natural Language Processing (NLP)* — 2023

- Led the large-scale deployment of a conversational AI agent for translating natural language queries into SQL (Text2SQL).
- Built an eval pipeline and benchmark dataset of 3,000+ real-world Text2SQL examples for robust evaluation.

Selected Research Experience

1. University of Illinois Urbana-Champaign (UIUC) | **Security & Privacy Group** [🌐] — Champaign, US

Graduate Researcher | *Advisor: Prof. Gang Wang* | *Areas: Machine Learning (ML) and Security & Privacy (S&P)* — 2022 - Present

- Identifying and protecting against practical underperformance, attacks, and misuse of AI models.
- Published projects at **IEEE S&P** and **USENIX Security**.

2. IIIT Delhi | **Laboratory for Computational Social Systems (LCS2)** [🌐] — Delhi, India

Undergraduate Researcher | *Advisor(s): Prof. Zubair Shafiq, Prof. Tanmoy Chakraborty* | *Areas: NLP, S&P and Graphs* — 2019 - 2021

- Identified malicious actors online using machine learning techniques.
- Published projects at **ACL** and **ICWSM**.

Teaching Experience

1. Introduction to Computer Science (CS124) | **UIUC** [🌐] — Champaign, US

Teaching Assistant | *Professor: Prof. Geoffrey Challen* — Fall 2022, Spring 2023, Fall 2023

- Served as Teaching Assistant for three consecutive semesters, engaging with 1,300+ students in each semester.
- Developed course materials, supervised quizzes, and held office hours.

2. Machine Learning Graduate (CS563) | **IIIT Delhi** — Delhi, India

Teaching Assistant | *Professor: Prof. Tanmoy Chakraborty* — Spring 2020

- Assisted teaching Graduate Machine Learning course to 180 students across undergraduate, master's, and doctoral levels.
- Conducted tutorials, resolved queries, graded quizzes, and developed the mid-semester examination.

Talks

Can LLMs identify Security Vulnerabilities? — April 2025

Advanced Computer Security, UIUC | *Slides*

Jailbreaking of LLMs — October 2024

LLM Post Pre-training, UIUC | *Slides*

Generating Sequences by Learning to Self-Correct — May 2024

Conversational AI, UIUC | *Slides*

Promptly Racing Ahead: A Survey on Multi-task Prompt-Based Learning — May 2023

Transfer Learning, UIUC | *Slides*

Editing Models with Task Arithmetic — April 2023

Transfer Learning, UIUC | *Slides*

A pseudo-labelling approach for low-resource NLP Domain Adaptation — December 2022

Trustworthy Machine Learning, UIUC | *Slides*

Weight Poisoning Attack on Pre-trained Language Model — July 2021

Security Machine Learning Seminar | *Slides*

Practical No-box Adversarial Attacks against DNNs — June 2021

Security Machine Learning Seminar | *Slides*

Academic Service

Reviewer	ACM COMPASS'21, ICLR'22 (Student Reviewer)
Volunteer	ACM FAccT'23, AAAI ICSWM'21, AAAI ICWSM'22, ACL'21