Nirav **Diwan**

CS PhD Student, University of Illinois Urbana-Champaign

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

Education

**Aug 2024 - **	**University of Illinois Urbana-Champaign (UIUC)** Ph.D in Computer Science Area: Security & Privacy, Machine Learning	**Champaign, US**
Aug 2022 May 2024	**University of Illinois Urbana-Champaign (UIUC)** Master of Science (M.S) in Computer Science (CS) (thesis-track) Specialization: Machine Learning	**Champaign, US**
Aug 2017 Jun 2021	**Indraprastha Institute of Information Technology, Delhi (IIITD)** Bachelor in Technology (B.Tech) in Computer Science & Engineering (CSE)	**Delhi, India**

Publications

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

[S.1] **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 [Under Review]* **[IEEE S&P'25]**

[C.3] **It Doesn't Look Like Anything to Me: Using Diffusion Model 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'24]**

[C.1] **Weakening the Inner Strength: Spotting Core Collusive Users in YouTube Blackmarket Network** [🔗] [GitHub]
Hridoy Sankar Dutta*, Nirav Diwan*, Tanmoy Chakraborty (* = Equal Contribution)
16^{th} *International AAAI Conference on Web and Social Media, 2022* **[ICWSM'22]**

[C.2] **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'21]**

[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 (Impact Factor = 4.2), 2020 **[Database'20]**

[W.2] **A Named Entity Based Approach to Modeling Recipes** [🔗] [GitHub]
Nirav Diwan, Devansh Batra and Ganesh Bagler
3^{rd} *International DECOR Workshop @ International Conference on Data Engineering, 2020* **[DECOR Workshop @ ICDE'20]**

[W.2] **Nutritional Profile Estimation in Cooking Recipes** [🔗] [GitHub]
Jushaan Singh Kalra, Devansh Batra, Nirav Diwan and Ganesh Bagler
3^{rd} *International DECOR Workshop @ International Conference on Data Engineering, 2020* **[DECOR Workshop @ ICDE'20]**

Research Experience

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

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

I am interested in practical adversarial attacks on Large Language Models (LLM's). Previously, I worked on developing robust data-driven approaches that address the underperformance and misuse of foundation models.

› *It Doesn't Look Like Anything to Me: Using Diffusion Model to Subvert Visual Phishing Detector*

Examined the threat of using foundation model-generated images for evading state-of-the-art phishing detectors.

- Developed a retrieval-based generative pipeline that can successfully evade existing phishing detectors.
- Conducted large-scale evaluation studies highlighting the effectiveness of the attack against several brands (e.g., Microsoft, Google).
- Contributed to the ideation, programming, and writing of the research paper.

Paper accepted at **USENIX Security Symposium 2024**.

› *You Can't Judge a Binary by Its Header: Data-Code Separation for Non-Standard ARM Binaries using Pseudo Labels:*

State-of-the-art methods cannot be adapted to perform binary analysis tasks on low-resource assembly language code. To mitigate this, we are developing a pseudo-label domain adaptation method using LLMs.

- Trained Large Language Models (LLMs) on a large-scale dataset of assembly code (20 million+ lines).
- Fine-tuned the LLM for downstream tasks and evaluated on several low-resource binary analysis tasks.
- Proposed a new evaluation metric for one of the fundamental problems of binary analysis - Code & Data Classification.

Under Review at **IEEE Security & Privacy Symposium, Oakland 2025**.

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. Collaborated with researchers from IIIT Delhi, IIT Delhi and University of California, Davis (UC Davis).

› *Fingerprinting Fine-tuned Language Models in the wild:*

Worked on the first empirical study on the risk posed by fine-tuned language models in generating convincing malicious text online using real-world data.

- Highlighted the threat of fine-tuned AI-generated text by comparing human-written and AI-generated text found online.
- Developed a classifier for identifying AI-generated text sources, achieving 87% precision in a 108-class classification task.
- Primary author of the paper, contributing to the ideation, programming and writing of the final paper.

Paper published in **ACL (Findings) 2021** conference as a primary author. Invited for a Live Q&A session at the RepL4NLP workshop. Invited for a poster presentation at Eastern European Machine Learning School (EEML) 2021.

› *Weakening the Inner Strength: Spotting Core Collusive Users in YouTube Blackmarket Network*

Led an investigation to identify anomalous users responsible for the growth of artificial appraisals on social media websites.

- Developed a graph-based deep learning framework for real-time anomaly detection in users.
- Proposed a modified core-periphery model for black markets. Conducted case studies demonstrating evasion tactics of users.
- Co-primary author, contributing to the ideation, programming and paper writing.

Paper published in **ICWSM 2022** conference as a co-primary author. Invited for a poster presentation at EEML 2022.

3. IIIT Delhi | **Insect Ecology, Evolution, and Conservation Lab** [🌐] Delhi, India

Undergraduate Researcher | *Advisor(s): Prof. Jainendra Shukla, Prof. Swapna Purandare* | *Areas: CV, Climate Science* 2019-2020

Contributed to projects aimed at improving the tracking of animals and insects in the wild.

› *Perceive the Pollinator*

Developed a Computer Vision (CV) model for identifying indigenous insects, aiding in their population tracking.

- Experimented with multiple state-of-the-art CV models including ResNet, EfficientNet.
- Achieved an F1-Score of 91.5% on a 5-class setting for the task of insect classification.
- Designed a framework for systematically labeling the images for the tasks of identification and classification.

4. Centre for Social Sciences and Humanities (CSH) | **French National Centre for Scientific Research** [🌐] Delhi, India

Summer Intern | *Advisor(s): Dr. Jean Thomas Martelli* | *Areas: NLP, Social Sciences* 2019

Served as a summer research intern, annotating and analyzing Indian Prime Ministers' speech data.

› *Populism in Indian Prime Minister's speeches*

Analyzed the speeches of Indian Prime Ministers to quantitatively assess populism exhibited by the leaders.

- Web crawled online datasets, contributing to the creation of speech dataset of Indian Prime Ministers.
- Performed a quantitative analysis of all Indian Prime Minister's speeches, helping to develop a relative scale of populism.

5. IIIT Delhi | **Complex Systems Lab** [🌐] Delhi, India

Research Intern | *Advisor: Prof. Ganesh Bagler* | *Areas: NLP, Information Extraction (IE), Databases* 2019-2020

Worked as a summer research intern to create a database of food resources online.

› *RecipeDB*

Contributed to the creation of RecipeDB [🌐] - an interactive database providing nutritional profiles of over 118,000 recipes. The website has over 5000+ monthly active users.

- Developed a Named Entity Recognition (NER) model for extracting ingredient info achieving an F1-Score of 95%.
- Scaled the model to process data from multiple data sources, processing over 118,000+ recipes.
- Conducted an exploratory data analysis of the recipe data (available on the website) on their diversity from across the world

RecipeDB was published in **Database: The Journal of Biological Databases and Curation (OUP) (Impact Factor = 5.8)**. The proposed NER model was accepted in the peer-reviewed **DECOR Workshop in ICDE conference**.

Teaching Experience

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

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

› Instructed introductory Java concepts to undergraduate students at UIUC.
› Served as a Teaching Assistant for three consecutive semesters, engaging with over 1,300 students.
› 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 in teaching the Graduate Machine Learning Course, which included 180 students from undergraduate to doctoral levels.
› Conducted tutorials, resolved student queries, graded quizzes, and developed the mid-semester examination.

Industrial Experience

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

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

- Worked on improving the safety and alignment of open-source LLMs for single-turn and multi-turn settings.
- Proposed and implemented sampling strategies that improved the performance on MT-Bench by a margin of 0.2-0.3 compared to state-of-the-art alignment techniques (e.g., DPO, ORPO, and SimPO).
- Conducted comprehensive evaluations on open-source LLM's (Mistral, LLAMA3, LLAMA3.1) on MT-Bench, TrueEval and QAEval.

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

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

- Led the scaling & deployment of a conversational AI agent that converts Natural Language Intents to SQL (NL2SQL) Commands.
- Designed a generative human-in–the-loop ML pipeline using Large Language Models (LLMs) to create a synthetic training dataset for NL2SQL, surpassing key performance metrics (KPM) by 5-10%.
- Developed Prompt Engineering Techniques using LLMs for NL2SQL. Increased precision, recall, and F1-Score by 10% over fine-tuning.
- Collected and annotated a dataset of over 3,000 samples for evaluating NL2SQL tasks in real-world settings

3. Prodigal Technologies | **ProVoice** [🌐] Remote, India

Machine Learning Engineer (MLE) | *Area: Natural Language Processing (NLP), Information Retrieval (IR)* 2021 - 2022

- Developed & deployed an Information Retrieval (IR) model for Event Identification in calls (F1-Score 93%) for 100,000 calls/day.
- Collaborated cross-functionally with engineering, product, and data teams to optimize search queries, improving efficiency by 10%.
- Managed 52 clients as a part of a three-member team of the ProVoice Team.

4. Research & Innovation Labs, Tata Consultancy Services (TCS) Research | **Networks and Graphs Team** [🌐] Delhi, India

Research Intern | *Area: Reinforcement Learning (RL)* 2020

- Designed time series forecasting methods for stock price prediction in Python using OpenAI Gym.
- Developed a data pipeline to collect and pre-process streaming data from up to 10 data sources using Python and Kafka.
- Simulated and tested a stock price prediction platform using Jenkins to perform over 100+ simulations.

Selected Projects

Instruct, Not Assist:LLM-based Multi-Turn Planning and Hierarchical Questioning for Socratic Code Debugging Feb'24 - Dec'24

Advisor: Prof. Dilek Hakkani Tur | *Areas: Natural Language Processing, Conversational AI*

- Introduced a novel approach leveraging Socratic questioning to improve students' critical-thinking skills for coding.
- Designed and implemented TreeInstruct, a state space-based planning algorithm that dynamically constructs personalized question trees to guide students in debugging tasks.
- Work currently under review at **EMNLP 2024**.

Leveraging Scene Graphs and Large Language Models for Visual Question Answering (VQA) Oct'23 - Dec'23

Advisor: Prof. Heng Ji | *Areas: Natural Language Processing, Computer Vision, Transfer Learning*

- Developed methods using scene graph-generated captions to enhance image compositionality understanding.
- Matched state-of-the-art performance for Image Captioning, Visual Question Answering (VQA), and Image-Text Captioning using text-only Language Models, as measured by standard benchmarks.

Promptly Racing Ahead: A Survey on Multi-Task Prompt Engineering of LLMs [🌐] Jan'23 - May'23

Advisor: Prof. Han Zhao | *Areas: Natural Language Processing, Transfer Learning*

- Conducted a comprehensive literature review on Multi-task Prompt Engineering for Large Language Models.
- Differentiated prompt-based learning into direct prompting and fine-tuning-based approaches, detailing their applications.
- Explored evaluation benchmarks and proposed future directions for prompt-based learning in LLMs, identifying gaps and opportunities for further research.

Pseudo-Label Domain Adaptation for LLM Fine-Tuning [🐙] Aug'22 - Dec'22

Advisor: Prof. Han Zhao | Areas: Trustworthy Machine Learning, Natural Language Processing

- Implemented the research paper Pseudo-Label Guided Unsupervised Domain Adaptation of Contextual Embeddings.
- Implemented Asymmetric Tri-training for Unsupervised Domain Adaptation, and achieved 75% F1-Score in sentence classification.

Talks

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*

Courses

- **Algorithms -** Data Structures and Algorithms, Algorithm Design and Analysis, Advanced Programming
- **Artificial Intelligence -** Advanced Natural Language Processing*, Transfer Learning, Trustworthy Machine Learning, Introduction to Data Mining, Deep Learning, Machine Learning, Fundamentals of Databases, Mining Large Networks, Big Data Analytics, Artificial Intelligence, Natural Language Processing, Applied Machine Learning (AML)*
- **Systems & Security -** Foundations of Computer Security, Cloud Computing and Distributed Systems, Computer Networks, Operating Systems, System Management, Computer Organisation
- **Mathematics -** Probability and Statistics, Multi-variable Calculus, Discrete Mathematics, Linear Algebra, Intro to Mathematical Logic
- **Other -** Social and Political Philosophy, Anthropology of Social Media, Environmental Science, Technical Communication, Portfolio Management, Introduction to Quantitative Biology, Body Language Studies

*Ongoing Coursework

Honours and Awards

Catalyzing Advocacy in Science and Engineering (CASE) Workshop 2024: Selected to represent UIUC at the Catalyzing Advocacy in Science and Engineering Workshop hosted by American Association for the Advancement of Science (AAAS) in Washington.

Google Computer Science Research Mentorship Scholarship (CSRMP) 2023 : 1 of 250 students globally awarded the scholarship.

Global Young Scientists Summit (GYSS) 2024 : 1 of 2 students selected to represent UIUC.

Dean's Thesis Appreciation Award, IIITD : Received thesis appreciation award for outstanding research from Dean of Academic Affairs.

Dean's Award for Academic Excellence, IIITD: Awarded to students with CGPA $>$ 9.0 for the academic years 2019-20 & 2020-21.

Summer Schools: Selected to present posters at Eastern European Machine Learning (EEML) 2021 and 2022, Oxford Machine Learning Summer School 2022, and Climate AI Summer School 2023.

Travel Grants: Received grants for virtual attendance and participation fees at *ACL 2021*, *AIIDE* 2020, *ICWSM* 2021 and 2022.

IIT-JEE Examinations: Secured All India Rank 1884 (top 0.15%) out of 1.5 million participants in the Indian Institutes of Technology Joint Entrance Examination (IIT-JEE) Mains 2017 and 4186 out of 200,000 participants in IIT-JEE Advanced 2017.

Leadership Roles

SecureML Seminar *Co-organizer* May'21 - Aug'21

- Co-organized the seminar and routinely presented research papers on Adversarial Machine Learning.

Computational Gastronomy Symposium (Season 3) *Co-organizer* Dec'19 - Jan'20

- Presented RecipeDB and organized the symposium alongside the research team and Dr. Ganesh Bagler. Attended by nationwide food specialists, anthropologists, and chefs.

Student Mentorship Organization, IIITD *Mentor* Aug'19 - May'21

- Mentored 7 Freshman B.Tech students for their academic and social growth in college for the academic year 2020-21. Continuing mentoring the students for graduate school.

Laboratory for Computational Social Systems (LCS2), IIITD *Participant* Aug'20 - May'21

- Routinely presented research papers at the weekly group meetings.

Shanti Sahyog *Volunteer* May'19 - Jun'19

- Volunteered at the NGO [🌍] Shanti Sahyog. Responsibilities included conducting a workshop on HTML, CSS, and Javascript for female high school students.

Academic Service

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