Mansi Phute

My research interests are **Responsible AI** and **ML safety**. I work on developing explanations for ML systems, analyzing them to identify vulnerabilities, and finding solutions to mitigate these issues. My UNDREAM system system offers a way to bridge differentiable rendering and photorealistic simulation for end-to-end adversarial attacks, thus enabling beter transferability of attacks to the physical world. My work includes LLM Self Defense, which leverages the model's own understanding of harm to protect itself from attacks.

mansiphute@gatech.edu
mphute.github.io
CV PDF
@mansiphute
@mphute
Google Scholar

Education

Summer 2024 — **Ph.D. in Computer Science**
Georgia Institute of Technology, Atlanta, GA

Fall 2022 — Spring 2024 **M.S. in Computer Science**
Georgia Institute of Technology, Atlanta, GA
Specialization: Machine Learning

Fall 2018 — Spring 2022 **B.Tech. in Electronics and Telecommunication**
Vishwakarma institute of Technology,
Honors: Artificial Intelligence and Data Analytics

Research Experience

Summer 2022 — Present **Georgia Institute of Technology**, Atlanta, GA
Graduate Research Assistant School of Computational Science and Engineering
Advisor: Duen Horng (Polo) Chau
Member of the Polo Club of Data Science where we bridge and innovate at the intersection of data mining and human-computer interaction to synthesize scalable, interactive, and interpretable tools that amplify human's ability to understand and interact with big data. Developed defences against adversarial attacks in Language and Vision domain.

Spring 2023 **Georgia Institute of Technology**, Atlanta, GA
Graduate Teaching Assistant School of Computational Science and Engineering
Mentor: Duen Horng (Polo) Chau

Fall 2021 — Spring 2022 **Nanyang Technological University**, Singapore
Undergraduate Research Assistant Cyber Security Research Centre at NTU (CYSREN)
Mentor: Thambipillai Srikanthan
Increasing python application security by analyzing libraries used. Developed dynamic dependency graph to trace vulnerabilities. Automated human resource planning and forecasting by combining business intelligence of NHS, UK with data analytics to properly shift the HR planning from manual to automated.

Spring 2021 **Vishwakarma Institute of Technology**, India
Undergraduate Research Assistant Associated with Dassault Systemes
Mentor: Jyoti Madake
Developed AI based solutions for agricultural problems faced in India by usimg hyperspectral imaging to predict soil fertility in the land

Fall 2020 **Vishwakarma Institute of Technology**, India
Undergraduate Research Assistant School of Electronics and Telecommunication
Mentor: Abha Marathe
Conducted a thorough literature survey on the use of AI in finance and the various ways it is used for risk management in the stock market

Industry Research Experience

Summer 2025 — Fall 2025 **HiddenLayer, Inc.**, Austin, TX

Research Assistant Adversarial Robustness Team
Mentor: Jason Martin, Ravi Balakrishnan
Helped pioneer transition of AI defense systems to account for multimodal attacks. Developed universal transferrable multimodal steering images that can alter model behavior using the input channel without requiring access to the model internals. My work during the internship was implemented into the product AIDR (AI Detection and Response) (https://www.hiddenlayer.com/aidr)

Summer 2019 **Tech Mahindra Ltd**, Pune, India
Intern, Web Development
Mentor: Rahul Bedmutha
Developed a portal for internal use, using HTML, CSS and Javascript.

Honors and Awards

2024 Marshall D. Williamson Fellowship
Awarded to a well-rounded, second-year Master's student who best embodies values of academic excellence and leadership

2022 Best Scholar in ECE
Merit-based award for the ECE student with the highest undergraduate GPA in the entire department

Publications

UNDREAM: Bridging Differentiable Rendering and Photorealistic Simulation for End-to-end Adversarial Attacks
Mansi Phute, Matthew Hull, Haoran Wang, Alec Helbling, ShengYun Peng, Willian Lunardi, Martin Andreoni, Wenke Lee, Duen Horng (Polo) Chau
arXiv. 2025.
Project PDF BibTeX

VISOR++ - Transferrable Visual Input based Steering for Output Redirection in Large Vision Language Models
Ravi Balakrishnan, **Mansi Phute**
arXiV. 2025.
Project PDF BibTeX

VISOR - Visual Input based Steering for Output Redirection in Large Vision Language Models
Mansi Phute, Ravi Balakrishnan
Assessing and Improving Reliability of Foundation Models in the Real World Workshop (AAAI). 2026.
Project PDF BibTeX

ComplicitSplat: Downstream Models are Vulnerable to Blackbox Attacks by 3D Gaussian Splat Camouflages
Matthew Hull, Haoyang Yang, Pratham Mehta, **Mansi Phute**, Aeree Cho, Haoran Wang, Matthew Lau, Wenke Lee, Willian Lunardi, Martin Andreoni, Duen Horng Chau
arXiv (arXiv). 2025.
Project PDF BibTeX

3D Gaussian Splat Vulnerabilities
Matthew Hull, Haoyang Yang, Pratham Mehta, **Mansi Phute**, Aeree Cho, Haoran Wang, Matthew Lau, Wenke Lee, Willian Lunardi, Martin Andreoni, Duen Horng Chau
CVPR Workshop on Neural Fields Beyond Conventional Cameras (NFBCC) (CVPR'25). 2025.
Project PDF BibTeX

Interpretation Meets Safety: A Survey on Interpretation Methods and Tools for Improving LLM Safety
Seongmin Lee, Aeree Cho, Grace C. Kim, ShengYun Peng, **Mansi Phute**, Duen Horng (Polo) Chau
Main, Conference on Empirical Methods in Natural Language Processing (EMNLP). 2025.
Project PDF BibTeX

RenderBender: A Survey on Adversarial Attacks Using Differentiable Rendering
Matthew Hull, Haoran Wang, Matthew Lau, Alec Helbling, **Mansi Phute**, Chao Zhang, Zsolt Kira, Willian Lunardi, Martin Andreoni, Wenke Lee, Duen Horng Chau
International Joint Conference on Artificial Intelligence (IJCAI) (IJCAI'25). 2025.
Project BibTeX

Semi Truths: A Large-Scale Dataset for Testing Robustness of AI-Generated Image Detectors
Anisha Pal, Julia Kruk, **Mansi Phute**, Manognya Bhattaram, Diyi Yang, Duen Horng (Polo) Chau, Judy Hoffman
NeurIPS. 2024.
Project PDF Code BibTeX

LLM Attributor: Interactive Visual Attribution for LLM Generation
Seongmin Lee, Zijie J. Wang, Aishwarya Chakravarthy, Alec Helbling, ShengYun Peng, **Mansi Phute**, Duen Horng (Polo) Chau,, Minsuk Kahng
ACL demo. 2024.
Project PDF BibTeX

Robust Principles: Architectural Design Principles for Adversarially Robust CNNs
ShengYun Peng, Weilin Xu, Cory Cornelius, Matthew Hull, Kevin Li, Rahul Duggal, **Mansi Phute**, Duen Horng (Polo) Chau, Jason Martin
BMVC. 2023.
Project PDF Code Video Poster BibTeX Best Poster Award

LLM Self Defense: By Self Examination, LLMs Know They Are Being Tricked
Mansi Phute, Alec Helbling, Matthew Hull, ShengYun Peng, Sebastian Szyller, Cory Cornelius, Duen Horng (Polo) Chau
ICLR Tiny Paper. 2024.
Project PDF Code BibTeX Deployed at ADP

Talks and Presentations

Large Language Model Evaluation
March 2025 Georgia Institute of Technology, CS 8001: Large Language Models
★ Invited

Large Language Models and How They Work
June 2024 Georgia Institute of Technology, CS 8001: Large Language Models
★ Invited

LLM Self Defense: By Self Examinations, LLMs Know They Are Being Tricked!
October 2023 IBM, San Jose CA

Press

April 2024 "Student Excellence Honored at Annual Event," Georgia Tech

August 2023 "GRE Success Stories: How Test Takers Scored Above the 90th Percentile," Jamboree

May 2020 "Team Eklavya- E&TC; students team Designs Autonomous sanitisation robot," Vishwakarma Institute of Technology

Teaching

Spring 2023 **Graduate Teaching Assistant**
Georgia Institute of Technology, Atlanta, GA
Data and Visual Analytics (CSE 6242 / CX 4242), Instructor: Duen Horng (Polo) Chau
I was a Teaching Assistant (TA) at Georgia Tech for the class Data and Visual Analytics where I worked with a team of 30 TAs to enable learning in a class of more than 1200 students. I was a part of designing homework and mentoring students in their course work and project work.

Fall 2019 **Teaching Volunteer**
Vishwakarma Institute of Technology, Pune, India
Aashadeep: Literacy Program for underprivileged people, Instructor:
A semester long teaching program where I created learning opportunities for increasing literacy in society aimed towards people outside the traditional schooling age. Thus proving that there is no binding of age to learn how to read or write. This program aimed at combating illiteracy in specific sections of society.

Service

Reviewer
NeurIPS Workshop on Socially Responsible Language Modelling (**NeurIPS SoLaR**) 2023
International Conference on Learning Representations (**ICLR**) 2026

Mentoring

Sri Ranganathan Palaniappan
B.S. in Computer Science, Georgia Institute of Technology

References

Dr. Polo Chau, Associate Professor
School of Computational Science and Engineering
Georgia Institute of Technology
cc.gatech.edu/~dchau/

Dr. Thambipillai Srikanthan, Professor
School of Computer Science and Engineering
Nanyang Institute of Technology
www.ntu.edu.sg/scse/about-us/past-chairs/prof-thambipillai-srikanthan