Mansi Phute

Mansi Phute (Adversarial ML + Explainable AI)

My research focuses on the **security** and **explainability** of language and vision foundation models. I work on developing explanations for machine learning models and analyzing them to identify vulnerabilities in existing ML systems, then find solutions to mitigate these issues. My work spans a wide range of application areas, including robust multi-object tracking in computer vision, developing defenses against attacks on large language models, and understanding large language models and the insights they can give us into human interactions.

I have collaborated with researchers, designers, and developers, at Intel Labs, Dassault Systems and Nanyang Technological University.

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

Education

Fall 2022 — Present
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
🏆 Highest GPA
🏆 Top 3 Best Graduating Students

Academic 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.

Fall 2021 — Spring 2022
Nanyang Technological University, Singapore
Undergraduate Research Assistant, Cyber Security Research Centre at NTU (CYSREN)
Mentor: Thambipillai Srikanthan
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 Systems
Mentor: Jyoti Madake
Developed AI based solutions for agricultural problems faced in India by usimg hyperspectral imaging to predict soil fertility in the land

Industry Experience

Summer 2019
Tech Mahindra Ltd, Pune, India
Intern, Web Development,
Mentor: Rahul Bedmutha
Developed AI based solutions for agricultural problems faced in India by usimg hyperspectral imaging to predict soil fertility in the land

Honors and Awards

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

2022
Best Student in ECE
One of the 3 students chosen as the best student in ECE department based on overall performance throughout undergrad

2020
2nd place in IEEE "One Million Seconds" Hackathon
Designed an autonomous system to support healthcare workers for cleaning the isolation wards in COVID-19 hospitals in India. First Runner Up from a total of 1200 participants

2015
Times NIE Student of the Year
Selected by "Times of India" as the best student in school based on overall acedamic and extracurricular performance.

Publications

All Publications

C5
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
British Machine Vision Conference (BMVC). 2023.
Project PDF Code BibTeX 🏆 #1 on RobustBench CIFAR-10 leaderboard , Best Poster

P4
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
Under review. .
Project PDF BibTeX

M3
Semantic Shift in Online Communities: Unmasking Information Flow through Linguistic Fingerprints
Julia Kruk*, Mansi Phute*, Amit Bhatacharjeee, Sanchita Porwal
Under Review. .
More coming soon! Project *Authors contributed equally

C2
A Survey on Machine Learning in Lithographys
Mansi Phute, Aditi Sahastrabudhe, Sameer Pimparkhede, Shubham Potphode, Kshitij Rengade, Swati Shilaskar
IEEE International Conference on Artificial Intelligence and Machine Vision. 2021.
Project BibTeX

C1
Precision Farming Based Soil Fertility Assessment Techniques
Mansi Phute, Jyoti Madake, Sripad Bhatlawande
IEEE International Conference on Advances in Computing, Communication, Embedded and Secure Systems. 2021.
Project BibTeX

Press

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 underpriviledged 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.

Mentoring

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

Service

Reviewer
NeurIPS Workshop on Socially Responsible Language Modelling (**NeurIPS SoLaR**) 2023

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

Contact

Mansi Phute `mansiphute@gatech.edu`
CODA Tech Square
Georgia Tech
756 W Peachtree St NW
Atlanta, GA 30308