



Mansi Phute

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 also worked with 3D simulation software like CARLA and Unreal Engine.

 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 Coßmputer 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 <i>Graduate Research Assistant School of Computational Science and Engineering</i> 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 <i>Graduate Teaching Assistant School of Computational Science and Engineering</i> Mentor: Duen Horng (Polo) Chau
Fall 2021 — Spring 2022	Nanyang Technological University , Singapore <i>Undergraduate Research Assistant Cyber Security Research Centre at NTU (CYSREN)</i> 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 <i>Undergraduate Research Assistant Cyber Security Research Centre at NTU (CYSREN)</i> 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 <i>Undergraduate Research Assistant Associated with Dassault Systems</i> Mentor: Jyoti Madake Developed AI based solutions for agricultural problems faced in India by using hyperspectral imaging to predict soil fertility in the land
Fall 2020	Vishwakarma Institute of Technology , India <i>Undergraduate Research Assistant School of Electronics and Telecommunication</i> 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
Summer 2019	Tech Mahindra Ltd , Pune, India <i>Intern, Web Development</i> 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
2023	Best Poster at BMVC'23 Robust Principles was awarded the Best Poster award at BMVC
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	Second 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

Publications

Semi Truths: A Large-Scale Dataset for Testing Robustness of AI-Generated Image Detectors Anisha Pal, Julia Kruk, Mansi Phute , Manogna Bhattaram, Diyi Yang, Duen Horng (Polo) Chau, Judy Hoffman <i>NeurIPS. 2024.</i>    
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 <i>ACL demo. 2024.</i>   
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 <i>BMVC. 2023.</i>       
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 <i>ICLR Tiny Paper. 2024.</i>     

Talks and Presentations

June 2024	Large Language Models and How They Work Georgia Institute of Technology, Course on Large Language Models ★ Invited
October 2023	LLM Self Defense: By Self Examinations, LLMs Know They Are Being Tricked! 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 <i>Georgia Institute of Technology, Atlanta, GA</i> 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 <i>Vishwakarma Institute of Technology, Pune, India</i> 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.

Grants and Funding

2023 — 2024	Guaranteeing AI Robustness against Deception Funded by DARPA: Defense Advance Research Projects Agency PIs:
-------------	--

Service

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

Mentoring

Sri Ranganathan Palaniappan <i>B.S. in Computer Science, Georgia Institute of Technology</i>
--

References

Dr. Polo Chau , Associate Professor School of Computational Science and Engineering <i>Georgia Institute of Technology</i> cc.gatech.edu/~dchau/
Dr. Thambipillai Srikanthan , Professor School of Computer Science and Engineering <i>Nanyang Institute of Technology</i> www.ntu.edu.sg/scse/about-us/past-chairs/prof-thambipillai-srikanthan