Poojitha Thota

PhD Candidate, The University of Texas at Arlington
Website: https://poojithathota.com/
Email: poojitha.thota@mavs.uta.edu
Contact: +18175503784

EDUCATION

- **2021 - 2026,** Ph.D. in Computer Science & Engineering, University of Texas at Arlington
  Dissertation: Systematic Approaches to Characterizing Vulnerabilities and Enhancing Robustness of Text and Vision-language Models
- **2019 - 2021,** MS Thesis in Computer Science & Engineering, University of Texas at Arlington
  Thesis: Analyzing Impact of Leaders' Statements on Spread of COVID-19
- **2014 - 2018,** B.Tech in Electronics & Communication Engineering, Vallurupalli Nageswara Rao Vignana Jyothi Institute of Engineering and Technology, Hyderabad, India

PUBLICATIONS

- **Poojitha Thota** and Shirin Nilizadeh. "Defending Text Summarization Models Against Data Poisoning", Submitted to *Usenix 2026*, Under review.
- **Poojitha Thota**, Tithy Tanusree, Faysal Shezan Hossain, and Shirin Nilizadeh. "The Insider's Advantage: Exploiting Automated Privacy Policy Analyzer Tools Through Subtle Text Manipulations", *ACM AsiaCCS*, *2026*.
- Seyyed Sadegh Moosavi, **Poojitha Thota**, Mohit Singhal, Abolfazl Asudeh, Gautam Das, Shirin Nilizadeh. "Unequal Privacy: Auditing Demographic Bias Vulnerabilities in Visual Protection Systems", *ACM AsiaCCS 2026*.
- Elham Pourabbas Vafa, Mohit Singhal, **Poojitha Thota**, Sayak Saha Roy, "Learning from Censored Experiences: Social Media Discussions around Censorship Circumvention Technologies", *2025 IEEE Symposium on Security and Privacy (SP)*.
- **Poojitha Thota** and Shirin Nilizadeh. "Attacks against Abstractive Text Summarization Models through Lead Bias and Influence Functions", *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, *2024*.
- Sayak Saha Roy, **Poojitha Thota**, Krishna Vamsi Naragam, and Shirin Nilizadeh. "From Chatbots to Phishbots?: Phishing Scam Generation in Commercial Large Language Models", *2024 IEEE Symposium on Security and Privacy (SP)*. **[Distinguished Paper Award]**
- **Poojitha Thota**, Jai Prakash Veerla, Partha Sai Guttikonda, Mohammad S. Nasr, Shirin Nilizadeh, and Jacob M. Luber. "Demonstration of an Adversarial Attack Against a Multimodal Vision Language Model for Pathology Imaging", *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*.
- Mohit Singhal, Chen Ling, Pujan Paudel, **Poojitha Thota**, Nihal Kumarswamy, Gianluca Stringhini, and Shirin Nilizadeh. "SoK: Content Moderation in Social Media, from Guidelines

to Enforcement, and Research to Practice", *2023 IEEE 8th European Symposium on Security and Privacy (EuroS&P)*.

- **Poojitha Thota** and Ramez Elmasri. "Web Scraping of COVID-19 News Stories to Create Datasets for Sentiment and Emotion Analysis", *The 14th Pervasive Technologies Related to Assistive Environments Conference (PETRA)*, *2021*.

WORK EXPERIENCE

UT Arlington - Teaching Assistant & Research Assistant in SPR Lab **Spring 2026**

- Assisting with instruction and grading for the course - Information Security.

UT Arlington - Research Assistant in SPR Lab **Fall 2025 - Current**

- Developed defenses against adversarial and data-poisoning attacks on LLMs, focusing on generative tasks such as text summarization across models (GPT-4, Gemini-2.5-Pro, Claude-3.5-Sonnet, Llama, Qwen-2.5,3, BART, T5, Pegasus, etc.).
- Built an automated framework revealing vulnerabilities in privacy-policy analysis tools using multi-level adversarial text perturbations, with ongoing work on improving system robustness.
- Curating multi-agent conversation framework as a defense against malicious LLM-based threats including jailbreaks, prompt injections, phishing attacks, content moderation, etc.
- Building an automated framework for evaluating adversarial robustness of continual learning mechanisms.

Google - Student Researcher Intern **Summer 2025**

- Contributed to Google Cloud Responsible AI team by curating a 10K+ adversarial prompt dataset across 11 jailbreak attack types to support classifier training on latest version of Gemini.
- Curated a stronger multimodal jailbreak benchmark by synthetically generating adversarial image–prompt pairs with Google's SIMULA, designed to remain effective against the latest models.

UT Arlington - Teaching Assistant in Department of CSE **Spring 2025**

- Assisted with instruction and grading for the course - Artificial Intelligence (AI).

Google - Student Researcher Intern **Fall 2024**

- Contributed to the Google Cloud Responsible AI team by enhancing safety features for Gemini.
- Investigated multi-agent conversation frameworks including Autogen, LangChain, OpenAI Swarm, and Google Onetwo, to improve safe prompt detection and optimize data filtration and annotation processes internally.
- Developed a safety filter to detect jailbreak prompts and devised a prompt embedding based approach that identifies unsafe inputs and refines them into safer versions.

UT Arlington - Research Assistant in SPR Lab **Fall 2022 - Summer 2024**

- Evaluated adversarial robustness of LLM- and VLM-based systems, including pathology-focused vision–language models (PLIP), identifying safety risks and model failure patterns.
- Developed pipelines to assess capabilities of commercial LLMs, including phishing website/email generation and detection of malicious user intent.

UT Arlington - Research Assistant in MAST Lab **Spring 2021 – Summer 2022**

- Worked for NSF funded project involving analysis of two sources of crowdsourced data to assess and build resilience in communities susceptible to natural disasters.
- Integrated with Streetwyze app to collect data from citizens and applying NLP techniques to understand the intensity of their statements.
- Integrated with StreetLight Data for recording location data, before and after a set of natural disasters, to monitor street level conditions and determine appropriate strategies to mitigate future disasters.

UT Arlington - Graduate Teaching Assistant in Department of CSE **Fall 2020 - Fall 2021**

- Assisted with instruction and grading for course - Theoretical concepts of CSE.

Hyundai Mobis - Graduate Engineer **July 2018 – July 2019**

- Developed Parking Application Software for Hyundai & Kia Genesis models, enabling remote driver access and optimizing WCET for parking distance logic. Worked in cross-functional teams using agile practices, including daily scrums and sprint planning.

TEACHING EXPERIENCE

- **CSE 4380/5380: Information Security -** Spring 2024, Spring 2026, Teaching Assistant.
- **CSE 4308/5360: Artificial Intelligence -** Spring 2025, Teaching Assistant.
- **RobustSumm: Exploring the Power and Pitfalls of Text Summarization Models -** Spring 2025, Designed and conducted workshop at OurCS@DFW, integrated with Student Computing Research Festival (SCRF).
- **CSE 3315, Theoretical concepts of CSE** - Fall 2020 - Fall 2021, Teaching Assistant, Guest Lecturer.

MEDIA INTERACTIONS

- "Research Spotlight - Attacks against abstractive text summarization models through lead bias and influence functions", covered by UT Arlington, December 2024.
- "AI: the drive to combat LLM-generated phishing attacks", interviewed by Hugo Sedouramane of Orange Research, August 2024.
- "UTA researchers work to prevent AI phishing scams", interviewed by Brian Lopez of Media Relations, UT Arlington, June 2024.

ACHIEVEMENTS

- **Outstanding Doctoral Dissertation Award (2026) at UTA:** Recognition for outstanding dissertation during the 2025-2026 academic year.
- **John S. Schuchman Outstanding Doctoral Student Award (2025) at UTA:** Recognition for outstanding research during the 2024-2025 academic year.
- **Best Runner-up Presentation Award (2025):** Awarded at the Student Computing Research Festival (SCRF) for workshop "RobustSumm: Exploring the Power and Pitfalls of Text Summarization Models".
- **Distinguished Paper Award:** Awarded at the 45th IEEE Symposium on Security and Privacy (S&P), 2024, for "From Chatbots to Phishbots?: Phishing Scam Generation in Commercial Large Language Models".
- **Student Travel Grant (2024):** Recieved for attending the 45th IEEE Symposium on Security and Privacy (S&P 2024).
- **Honorable Presentation Award (2024):** Awarded at the Student Computing Research Festival (SCRF) for "Attacks against abstractive text summarization models through lead bias and influence functions".

SERVICE

Reviewer ACL, EMNLP, NAACL	2024 - Present
Reviewer CVPR (AFDM Workshop)	2026
PC Member CSET (ACSAC Workshop)	2025
Sub-Reviewer USENIX Security Symposium	2024
Sub-Reviewer NDSS	2023-2025
Sub-Reviewer CCS	2024, 2026
Sub-Reviewer RAID Symposium	2023-2024
Doctoral Student Mentor (UT Arlington)	2024 -Present

REFERENCES

- Dr. Shirin Nilizadeh, Associate Professor, Department of Computer Science, UT Arlington.
- Dr. Jacob Luber, Associate Professor, St. Jude Children's Research Hospital.
- Dr. Faysal Shezan, Assistant Professor, Department of Computer Science, UT Arlington.