POOJITHA THOTA

+1(817) 550-3784, Dallas, TX | poojitha.thota@mavs.uta.edu | https://poojithathota.com/

EDUCATION

PhD Candidate in Computer Science and Engineering, University of Texas at Arlington **Aug 2021- May 2026 (Expected)**
Master's in computer science and engineering, University of Texas at Arlington **Aug 2019- May 2021**
Bachelor of Technology in Electronics & Communication, VNRVJIET, India **Aug 2014- May 2018**

RESEARCH INTERESTS

NLP, LLM Safety & Robustness, Adversarial ML, Multimodal/Vision Language Models, Security & Privacy, Content Moderation, Data Mining, Social Network Analysis, Multi-Agent Systems.

TECHNICAL SKILLS

Programming Languages	: Python, R, C, Java.
AI/ML & LLM Frameworks	: Pytorch, TensorFlow, Scikit-learn, Keras, LangChain, OpenCV, NLTK
LLM & GenAI Tooling	: RAG pipelines, multi-agent frameworks (Autogen, LangChain, OpenAI Swarm).
Cloud & Infrastructure	: AWS, Google Cloud Platform (GCP), REST APIs.
Web Technologies	: HTML5, CSS, JavaScript, Bootstrap, Node JS, React.js, React Native, Laravel.
Databases	: SQL, MySQL, MongoDB.
Operating Systems	: Linux, Windows, Mac OS.

PUBLISHED PAPERS

- Sadegh Moosavi Khorzooghi, **Poojitha Thota**, Mohit Singhal, Abolfazl Asudeh, Gautam Das, Shirin Nilizadeh "Unequal Privacy: Auditing Demographic Bias Vulnerabilities in Visual Protection Systems", *AsiaCCS 2026*.
- Elham Pourabbas Vafa, Mohit Singhal, **Poojitha Thota**, Sayak Saha Roy, "Learning from Censored Experiences: Social Media Discussions around Censorship Circumvention Technologies", *IEEE S&P 2025*.
- **Poojitha Thota**, and Shirin Nilizadeh, "Attacks against Abstractive Text Summarization Models through Lead Bias and Influence Functions", *EMNLP 2024.*
- Roy, Sayak Saha, **Poojitha Thota**, Krishna Vamsi Naragam, and Shirin Nilizadeh. "From Chatbots to Phishbots?: Phishing Scam Generation in Commercial Large Language Models", *IEEE S&P 2024, **"Distinguished Paper Award"***.
- **Thota, Poojitha**, Jai Prakash V, Partha Sai G, Mohammad S. Nasr, Shirin Nilizadeh, and Jacob M. Luber. "Demonstration of an Adversarial Attack Against a Multimodal Vision Language Model for Pathology Imaging", *IEEE ISBI 2024.*
- Singhal, Mohit, 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", *EuroS&P 2023*.
- **Thota, Poojitha,** and Elmasri Ramez. "Web scraping of covid-19 news stories to create datasets for sentiment and emotion analysis", *PETRA 2021.*

EXPERIENCE

UT Arlington – Research Assistant in Security & Privacy Research Lab **Aug 2025 - Current , Aug 2022 – Aug 2024**

- 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.).
- Evaluated adversarial robustness of LLM- and VLM-based systems, including pathology-focused vision–language models (PLIP), identifying safety risks and model failure patterns.
- Built an automated framework revealing vulnerabilities in privacy-policy analysis tools using multi-level adversarial text perturbations, with ongoing work on improving system robustness.
- Developed pipelines to assess capabilities of commercial LLMs, including phishing website/email generation and detection of malicious user intent.

Google- Student Researcher Intern **May 2025 – Aug 2025**

- Designed and operationalized a 10K+ adversarial prompt pipeline across 11 jailbreak classes to support large-scale classifier training and safety evaluation for 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 **Jan 2025 – May 2025**

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

Google - Student Researcher Intern **Aug 2024 – Jan 2025**

- Contributed to the Google Cloud Responsible AI team by enhancing safety features for Gemini. Prototyped multi-agent safety architectures (Autogen, LangChain, OpenAI Swarm, Google OneTwo) to improve unsafe prompt detection and routing in production workflows.

- 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 Mining and Analysis in Data lab **Dec 2021 – Aug 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 **Aug 2020 – Dec 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.

PROJECTS

Exploiting Automated Privacy Policy Analyzer Tools Through Subtle Text Manipulations **Submitted to USENIX 2026**

- Exposed vulnerabilities in automated policy analyzers by developing the APATRA framework, showing that subtle policy text perturbations can bypass state-of-the-art LLMs and policy tools. Paper under review.

Defending Text Summarization Models via Machine Unlearning and Poisoning Detection **Submitted to IEEE S&P 2026**

- Proposed the first defense framework for text summarization models using machine unlearning, mitigating lead bias and detecting/removing data poisoning with up to 90% recovery and minimal performance loss. Paper under review.

Attacks against Text Summarization Models through Lead Bias and Influence Functions **EMNLP 2024**

- Investigated vulnerabilities in Text Summarization Models through Adversarial Perturbations and Data Poisoning Attacks.
- Identified and exploited inherent biases to perform adversarial perturbations and utilized influence functions to poison Text Summarization datasets effectively. Targeted LLMs include BART, T5 and Pegasus.

From Chatbots to PhishBots? - Preventing Phishing scam generation in commercial LLMs **IEEE S&P 2024**

- Demonstrated that commercial LLMs can generate evasive phishing attacks through crafted malicious prompts, with results published at IEEE Symposium on Security and Privacy 2024 and acquired "Distinguished Paper Award".
- Developed a classifier achieving 98% accuracy for early detection of malicious prompts from user conversation history, available at https://huggingface.co/phishbot/ScamLLM. Released as a ChatGPT Plugin.

Adversarial Attack on Multimodal Vision Language Model for Pathology Imaging **IEEE ISBI 2024**

- Evaluated and identified vulnerabilities in pre-trained VLMs, including PLIP, for Pathology Imaging.
- Implemented Projected Gradient Descent (PGD) attacks with100% success rate in revealing weaknesses, with findings published and presented at IEEE International Symposium on Biomedical Imaging (ISBI), 2024.

Master's Thesis - Analyzing impact of Leaders statements on spread of COVID-19 **Spring 2021**

- Analyzed emotion and sentiment in leaders' statements and news stories using BERT, integrating results with COVID-19 case data from JHU-CSE to achieve 85.2% accuracy.
- Built a new dataset by scraping federal/state officials' statements and major news outlets (CNN, Washington Post, BBC).

Driver Drowsiness and Distraction Alert System **Fall 2022**

- Developed a driver-assist system that integrates face recognition (via OpenCV detecting eyelid and mouth movements) and speech recognition for interactive command processing. Integrated our system with Spotify and Google to enable music playback and to locate nearby rest areas or cafes.

Rico Hierarchy to React Native **Fall 2021**

- Developed a web-based tool to automatically convert UI/UX JSON datasets into React Native code. Achieved 100% accuracy in mapping UI components (text, images, buttons) and ensured compatibility across Android and iOS platforms.

Blockchain and Web Development **Spring - Summer 2020**

- Developed a dynamic shopping website using HTML5, CSS, JavaScript, PHP, Bootstrap, and MySQL, enabling bitcoin transactions via REST API calls. Integrated the web application with Laravel.