POOJITHA THOTA

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

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

Security and Privacy, Natural Language Processing, Adversarial Machine Learning, Multimodal Learning, Large Language Models, Vision Language Models, Social Network Analysis, Content Moderation, Data Mining

TECHNICAL SKILLS

Programming Languages	: Python, R, C, Java
Web Technologies	: HTML5, CSS, JavaScript, Bootstrap, Node JS, React.js, React Native, Laravel
Database & Cloud Technologies	: SQL, MySQL, MongoDB
Machine Learning Frameworks	: Keras, PyTorch, Pandas, Tensorflow, NumPy, Scikit-learn, NLTK, OpenCV
Operating Systems	: Linux, Windows, Mac OS

PUBLISHED PAPERS

- **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", In *2024 IEEE Symposium on Security and Privacy (SP)*
- **Thota, Poojitha**, 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", In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*
- 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", In *2023 IEEE 8th European Symposium on Security and Privacy (EuroS&P)*
- **Thota, Poojitha,** and Elmasri Ramez. "Web scraping of covid-19 news stories to create datasets for sentiment and emotion analysis", In *The 14th pervasive technologies related to assistive environments conference*

EXPERIENCE

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 – Research Assistant in Security and Privacy Lab — **Spring 2025**

- Worked on designing robust defenses for the adversarial attacks and data poisoning attacks against large language models, performing regenerative tasks including text summarization.
- Developed a novel framework to expose vulnerabilities in automated privacy policy analysis systems using multi-level adversarial text perturbations, with ongoing research into robust defense strategies.

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 Security and Privacy Lab — **Fall 2022 – Summer 2024**

- Explored the adversarial robustness of LLMs performing text summarization, and VLMs utilized for medical domain. Targeted text summarization models include GPT-3.5, Gemini-1.5-Pro, Claude-3.5-Sonet, BART, T5 and Pegasus; and VLMs include PLIP for Pathology Imaging.
- Worked on evaluating the capabilities of commercial LLMs to create phishing websites and emails and implemented methods to detect malicious intent of users.

UT Arlington - Research Assistant in Mining and Analysis in Data lab — **Dec 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 **Aug 2020 – Dec 2021**

- Grader cum teaching assistant 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 during parking and optimizing WCET for parking distance logic.
- Collaborated in cross-functional teams using agile methodology with 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.

Defending Text Summarization Models via Machine Unlearning and Poisoning Detection **Submitted to NDSS 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.

Multi-Agent Framework for Conversational Safety **Submitted to EMNLP 2025**

- Introduced a novel multi-agent content moderation framework where specialized agents collaboratively verify safety, achieving 4–11% higher accuracy and lower false positives than single-model and state-of-the-art moderation systems

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 awarded "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) and applying NLP techniques such as POS tagging, chunking, and named entity recognition.

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.
- *Skillset used:* OpenCV, Google Cloud API, Spotify Web API, Python

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.
- *Skillset used:* JSON, React Native, Python

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.
- *Skillset used:* Matic Testnet, Komodo RPC, Apache Tomcat, Xampp, and Web Technologies, HTML5, CSS, JS, Bootstrap, PHP and MySQL