Aravind Cheruvu

Portfolio Linkedin Google Scholar aravindcheruvu2024@gmail.com ☎ +1-540-824-8618

Education

Virginia Tech Blacksburg, Virginia
Ph.D. in Computer Science (+MS): GPA: 3.75/4.00 Advisor:* ***Bimal Viswanath*** *Aug. 2021 – Present*

Jawaharlal Nehru Technological University Hyderabad, India
Bachelor of Technology in Information Technology: GPA: 8.51/10.0 *Aug. 2012 – May. 2016*

Selected Projects

Framework for Mitigating Toxicity while Customizing Conversational AI *Under submission*

- Proposed a novel plug-and-play defense framework aimed at mitigating the toxicity during chatbot customization.
- Evaluated 2 LLM-based toxicity classification approaches on 16 LLMs, demonstrating their superior performance of 25% over industry API services and significant reduction in toxicity under a data poisoning setting.
- Designed a synthetic data generation approach to reinforce desired conversational behavior and align models through data-level intervention and direct preference optimization (DPO), significantly reducing toxicity to $\sim 0\%$ for chatbots.

Toxicity Injection Attacks on Open-domain Chatbots *Published in ACSAC'23*

- Conducted a study on the susceptibility of various chatbots to toxicity injection attacks in Dialog-based learning (DBL)
- Highlighted the risk of adversaries altering chatbot toxicity levels and devising backdoor attack strategies using LLMs
- Demonstrated the limitations of existing defenses against adaptive attacks using LLM-based software agents, where toxicity exhibited was $\sim 18\%$ even after applying the best defenses.

System and Method to Generate Time-Profiled Temporal Pattern Tree *Indian Patent No. 397728*

- Invented and patented a cost-efficient temporal tree structure for storing time-series transactional data, implementing an algorithm to reveal interesting patterns that reduces execution time by 90% and memory utilization by 80%.

Selected Publications

Submitted to IEEE S&P'25 *Title anonymized for double blind submission* **1st** author

Submitted to IEEE S&P'25 *Title anonymized for double blind submission* **2nd** author

IEEE S&P'24 *Analysis of Recent Advances in Deepfake Image Detection in an Evolving Threat Landscape* **2nd** author

ACSAC'23 *A First Look at Toxicity Injection Attacks on Open-domain Chatbots* **1st** author

IADIS'17 *Journal: A dissimilarity measure for mining similar temporal association patterns* **4th** author

ICEMIS'17 *Using normal distribution to retrieve temporal associations by Euclidean distance* **1st** author

ICEMIS'16 *Estimating temporal pattern bounds using negative support computations* **1st** author

Technical Skills

GenAI technologies: LLMs, Model customization, LoRA Fine-tuning, Safety alignment, SFT, DPO, Stable diffusion, StyleGAN, Deepfakes generation and detection

Machine learning framework / libraries: Huggingface Transformers, Tokenizers, PEFT, Accelerate, DeepSpeed, PyTorch, Numpy, Scikit-Learn, Pandas, TRL, Tensorflow

Programming Languages: Python, Java, C, C++, HTML/CSS

Developer Tools: Oracle SQL, Linux, GitHub, SQL Developer, VS Code, Eclipse, Netbeans, Android Studio, Weka

Experience

Virginia Tech, Graduate Research Assistant Dec. 2021 - present

- Led research on conversational AI building Responsible AI systems, particularly chatbots, with a focus on investigating and mitigating toxicity in chatbots and model customization pipelines. Exploring attacks and defenses using state-of-the-art Large Language Models (LLMs).
- Specialized in deepfakes, GANs, and diffusion models within the CV domain. Conducted large-scale evaluations of deepfake detector robustness, identifying critical vulnerabilities and improving detection systems.

Deloitte Consulting, Senior Consultant ← Consultant ← Analyst Dec. 2016 - Jul. 2021

- **Certified Oracle HCM Cloud transformation consultant with 4.5 years of experience:** Designed 50+ Technical RICEF objects, performed fit-gap analysis, and led teams and performed $MM Payroll data analysis for 5 large-scale US client implementations, identifying, mitigating system defects and efficiently communicating cost and operational impacts.