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:* ***Dr. Danfeng Yao*** — *Aug. 2021 – May. 2026**

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*

- Developed **TuneShield**, a scalable and robust defense framework to mitigate toxicity during LLM fine-tuning, enabling safe and user-tailored chatbot customization.
- Designed **LLM-based toxicity classification**, **synthetic healing data** generation, and **DPO-based alignment** methods to filter toxic content and reinforce desired conversational behaviors, achieving $\sim 0\%$ toxicity while preserving model utility. The framework exhibits resilience even in biased classifier and adversarial attack scenarios.

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

- Conducted an extensive study on the vulnerability of chatbots to **toxicity injection and backdoor attacks** in dialog-based learning (DBL).
- Evaluated state-of-the-art defense methods against adaptive LLM-based attack agents, revealing residual toxicity levels of approximately $\sim 18\%$ and exposing critical gaps in existing safety frameworks.

Selected PhD Publications

In Review *TuneShield: Mitigating Toxicity in Conversational AI while Fine-tuning on Untrusted Data* $\mathbf{1^{st}}$ author
In Review *Taming Data Challenges in ML-based Security Tasks: Lessons from Integrating Generative AI* $\mathbf{2^{nd}}$ author
IEEE S&P'24 *Analysis of Recent Advances in Deepfake Image Detection in an Evolving Threat Landscape* $\mathbf{2^{nd}}$ author
ACSAC'23 *A First Look at Toxicity Injection Attacks on Open-domain Chatbots* $\mathbf{1^{st}}$ 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

Samsung Research America (SRA), GenAI Research Intern — Aug. 2025 – Nov. 2025

- Designed and developed Generative AI applications for digital health and wellness, to deliver adaptive coaching, personalized recommendations, and context-aware health insights.
- Developed scalable backend pipelines and **RESTful APIs** to process multi-modal health data from smartphones and wearables, enabling **RAG-based AI assistants** that provide real-time, evidence-informed guidance.
- Partnered with **AI scientists, clinicians, and human factors researchers** to co-innovate digital health solutions. Implemented AI-driven insights and explainable visualizations that translated wearable and time-series analytics into actionable wellness feedback. Supported pilot studies to validate GenAI-based health coaching and conversational frameworks.

Virginia Tech, Graduate Research Assistant — Dec. 2021 - present

- Led research on conversational AI building Responsible AI systems, 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.