Aravind Cheruvu

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

EDUCATION

Virginia Tech Blacksburg, Virginia
Ph.D. in Computer Science: GPA: 3.77/4.00 Advisor:* ***Dr. Danfeng Yao*** *Aug. 2021 – June. 2026**

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

SELECTED PROJECTS

Defense Framework for Mitigating Harmful Fine-tuning Attacks in Conversational AI *In Submission*

- Developed SafetyOpus, a defense framework for mitigating harmful fine-tuning attacks during LLM chatbot customization, covering a broad taxonomy of 14 harm categories (e.g., hate speech, self-harm, drug abuse, child abuse) beyond conventional toxicity.
- Designed a two-stage pipeline integrating LLM-based guardrail filtering (LlamaGuard, Qwen3Guard, WildGuard) with synthetic healing data generation and preference optimization, instantiated as two-step DPO and single-step ORPO variants, restoring harmlessness rates to ~ 1.0 and counteracting alignment drift without external alignment datasets.
- Validated robustness across 3 chatbots (LLaMA, Mistral, Qwen), achieving the safety goal in 8 of 9 cases with $\sim 14.1\%$ harmlessness improvement on unsafe contexts; demonstrated generalization beyond imperfect filters (e.g., LlamaGuard at 55% recall) and outperformed the state-of-the-art StarDSS defense baseline.

Framework for Mitigating Toxicity while Customizing Conversational AI *Accepted at CODASPY'26*

- Developed TuneShield, a Responsible AI framework for safe LLM fine-tuning. Orchestrated AI-based workflow for synthetic healing data generation to autonomously mitigate toxicity while enabling user-tailored customization.
- Designed LLM-based toxicity classification and DPO-based alignment methods to filter toxic content and reinforce desired conversational behaviors, achieving $\sim 0\%$ toxicity (from 30% injection rates) while preserving model utility.
- Validated defense robustness against adaptive adversarial and jailbreak attacks, utilizing PromptAttack and optimization-based tools like AmpleGCG-Plus, ensuring resilience even when safety classifiers were compromised.

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

- Simulated Agentic AI threats by engineering automated LLM-based malicious agents that masquerade as benign users to inject toxicity and backdoors into chatbots during 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.

Adversarial Deepfakes via Vision Foundation Models *Published in IEEE S&P 2024*

- Engineered a "noise-free" evasion attack by adversarially updating StyleGAN generator weights using Vision Foundation Models (e.g., CLIP, ViT) as surrogate classifiers, embedding evasion patterns directly into the generation process.
- Demonstrated that VFM-guided semantic manipulation effectively bypasses detection, degrading the recall of 8 state-of-the-art detectors by up to 88% and revealing critical vulnerabilities in defenses against adaptive generative adversaries.

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

- Designed and patented a compute-efficient temporal tree structure for time-series data, implementing a novel similarity-profiled association rule mining algorithm that reduces execution time by 90% and memory utilization by 80% compared to naive approaches.

SELECTED PHD PUBLICATIONS

CODASPY'26 Optimus: A Robust Defense Framework for Mitigating Toxicity while Fine-Tuning Conversational AI **1st** author

ASIA CCS'26 Taming Data Challenges in ML-based Security Tasks: Lessons from Integrating Gen AI **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

TECHNICAL SKILLS

GenAI & Agentic AI: LLMs, Autonomous Agents, RAG, LangChain, LlamaIndex, Model Customization, PEFT (LoRA/QLoRA), SFT, DPO, RLHF, Chain-of-Thought (CoT), Stable Diffusion, StyleGAN, Deepfakes

AI Safety & Adversarial Security: Responsible AI, Safety Alignment, Red Teaming, Prompt Injection, Jailbreak Attacks (GCG, AmpleGCG-Plus), Backdoor Attacks, Data Poisoning, Adversarial Perturbations (TextFooler, PromptAttack), Toxicity Mitigation

Machine Learning Frameworks: HuggingFace (Transformers, TRL, PEFT, Accelerate), DeepSpeed, PyTorch, TensorFlow, NumPy, Scikit-Learn, Pandas

Programming & Developer Tools: Python, Java, C, C++, HTML/CSS, SQL, Linux, Git/GitHub, VS Code, Docker

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 validating GenAI 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 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.

Tata Consultancy Services, Assistant System Engineer - Trainee Jun. 2016 - Sep. 2016

- Trained in E-Business Suite, Oracle Business Intelligence EE and Oracle Data Integrator tools.

ACHIEVEMENTS AND MEDIA COVERAGE

- **Pratt Fellowship**, Department of Computer Science, Virginia Tech
- **CCI SWVA Cyber Innovation Scholarship**, for FY '23 and '24
- **Best Poster Award at CCI Researcher Showcase**, Virginia Tech 2023
- **News Interview at WDBJ7**, "Virginia Tech research aims to reduce toxic language from artificial intelligence."
- **News interview to VPM News Focal Point**, "Artificial intelligence: What are the risks and benefits?"
- **Gold Medal for best outgoing student**, Department of I.T (Bachelors)

PROFESSIONAL SERVICE

- **Reviewer** for Journal: Expert Systems With Applications
- **Reviewer** for Journal: IEEE Transactions on Dependable and Secure Computing
- **Program committee** for ACL 2026 Industry Track