Ali **Derakhshan**

Research Assistant · Teaching Assistant ·

Department of Computer Science, Donald Bren School of ICS, University of California, Irvine, 3054 Donald Bren Hall, CA 92697, USA

aderakh1@uci.edu | aliderakhsh | aliderakhsh | Ali Derakhshan

Personal Profile

I am a dedicated Ph.D. candidate at the University of California, Irvine, specializing in trustworthy machine learning and AI, with extensive experience in adversarial defenses for LLMs, social engineering detection, and multimodal systems. My interdisciplinary projects—supported by NSF funding—have garnered over 200 citations, demonstrating my ability to produce high-impact research. In addition to my strong academic record (including a near-perfect M.Sc. GPA), I bring integrity, a collaborative spirit, and a passion for advancing AI safety. I actively review for reputable journals, reflecting my commitment to research excellence and the broader scientific community.

Education

University of California, Irvine — *Irvine, USA*

Ph.D. Candidate in Computer Science — *Since Sep. 2019*

- Research Focus: ML and AI Trustworthiness in NLP, LLMs, VLMs, and Multi-Modal Models, emphasizing alignment, safety, and reliability.

University of California, Irvine — *Irvine, USA*

M.Sc. in Computer Science, **GPA: 3.98/4.0** — *Mar 2023*

- Completed course-based Master's degree during Ph.D.

Sharif University of Technology — *Tehran, Iran*

M.Sc. in Computer Engineering, **GPA: 3.9/4.0** — *Sep 2014 – Feb 2017*

- Specialization: Artificial Intelligence and Robotics
- Thesis: "Analyzing Purchase Satisfaction Using Opinion Mining", Advisor: Prof. Hamid Beigy

K.N. Toosi University of Technology — *Tehran, Iran*

B.Sc. in Computer Engineering - Hardware, **GPA: 3.52/4.0** — *2009 – 2014*

- Thesis: "Text Summarization Using LSA and NMF"

Computer & Programming Skills

Languages	Persian (Native), English (Proficient)
Programming	Python, SQL, MATLAB, Java, C/C++
Software	NumPy, SciPy, Matplotlib, Pandas, Sklearn, Pytorch, TensorFlow, Huggingface
Document Creation	LaTeX, Markdown, Adobe Photoshop, Adobe Premiere Pro

Experience

Research Assistant — *Prof. Ian Harris*

Secure Systems and Software Laboratory, University of California, Irvine — *Since Sep. 2019*

- **Adversarial Prompt Shield (APS):** Developed a robust safety system to defend against jailbreaking attacks on Large Language Models (LLMs), enhancing AI trustworthiness, alignment, safety, and reliability. Engineered gradient-based techniques to generate adversarial suffixes and introduced adversarial training datasets to improve model resilience. Fine-tuned LLMs to enhance performance and adaptability. Published results in the Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH 2024) at the North American Chapter of the Association for Computational Linguistics (NAACL).
- **Detection of Telephone-Based Social Engineering Attacks:** Created novel machine learning approaches for detecting social engineering attacks based on semantic content and speech acts, contributing to AI safety in cybersecurity.
- **Cyberbullying and Misogyny Detection:** Applied advanced NLP techniques and transfer learning with Compact BERT models to detect hate speech, cyberbullying, and misogynous content, focusing on AI trustworthiness in NLP and multi-modal models.
- **Pioneering Human Studies on Telephone Scams:** Led groundbreaking research involving 186 participants in simulated telephone scam scenarios, uncovering critical insights into human susceptibility to social engineering attacks. Developed a unique dataset of recorded scam calls, contributing vital resources to the cybersecurity research community. Published findings in the International Workshop on Socio-Technical Aspects in Security and Trust (Springer, 2020).
- **NSF-Funded Research:** Conducted Ph.D. research under the NSF grant "Detecting Social Engineering Attacks Using Semantic Language Analysis" (Award ID: 1813858), advancing the field of AI-driven cybersecurity and social engineering detection.

Research Assistant — *Prof. Hamid Beigy*

Sharif University of Technology, Tehran, Iran — *2017 - 2019*

- **Opinion Mining for Customer Satisfaction:** Conducted innovative research on opinion mining techniques to analyze customer purchase satisfaction and its impact on stock market fluctuations.
- **Large-Scale Social Media Analysis:** Leveraged advanced machine learning and probabilistic inference techniques to process and analyze extensive customer opinions from various social media platforms.
- **Master's Thesis:** Completed thesis titled "Analysing purchase satisfaction using opinion mining," contributing to the field of sentiment analysis and its applications in market research.

Data Scientist Intern

Digikala *Summers 2014 & 2016*

- **Graph Data Visualization:** Developed innovative techniques for visualizing large-scale networks, implementing multi-cluster pie charts to effectively represent complex node relationships and connections.
- **Customer Sentiment Analysis:** Conducted analysis on purchase satisfaction using comment data, contributing to improved user experience and product recommendations.

Teaching Assistant

Department of Computer Science, University of California, Irvine *Since Sep 2021*

Graduate Courses:

- CS 273A: Machine Learning (Prof. Alexander Thomas Ihler): Fall 2022, Fall 2024
- CS 274P: Neural Networks and Deep Learning (Prof. Pierre Baldi): 2023
- CS 244: Intro to Embedded Systems (Prof. Bozorgzadeh): Spring 2024
- CS 273A: Machine Learning (Prof. Stephan Mandt): 2020

Undergraduate Courses:

- ICS 31: Intro to Programming (Prof. Ian Harris): 2022
- CS 145: Embedded Software (Prof. Ian Harris): 2021

Publications

Jinhwa Kim, **Ali Derakhshan**, Ian Harris

[1] Robust Safety Classifier Against Jailbreaking Attacks: Adversarial Prompt Shield *2024 - Published*

☑ Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH 2024) at NAACL

Jinhwa Kim, **Ali Derakhshan**, Ian Harris

[2] Robust safety classifier for large language models: Adversarial prompt shield *2023 - Preprint*

☑ arXiv preprint arXiv:2311.00172

Mitra Behzadi, **Ali Derakhshan**, Ian Harris

[3] Mitra behzadi at semeval-2022 task 5: Multimedia automatic misogyny identification method based on CLIP *2022 - Published*

☑ Proceedings of the 16th International Workshop on Semantic Evaluation

Ali Derakhshan, Ian Harris, Mitra Behzadi

[4] Detecting telephone-based social engineering attacks using scam signatures *2021 - Published*

☑ Proceedings of the 2021 ACM Workshop on Security and Privacy Analytics

Mitra Behzadi, Ian Harris, **Ali Derakhshan**

[5] Rapid Cyber-bullying detection method using Compact BERT Models *2021 - Published*

☑ 2021 IEEE 15th International Conference on Semantic Computing (ICSC)

Ian G Harris, **Ali Derakhshan**, Marcel Carlsson

[6] A Study of Targeted Telephone Scams Involving Live Attackers *2020 - Published*

☑ International Workshop on Socio-Technical Aspects in Security and Trust

Ali Derakhshan, Hamid Beigy

[7] Sentiment analysis on stock social media for stock price movement prediction *2019 - Published*

☑ Engineering applications of artificial intelligence

Services

TDSC Reviewer

Transactions on Dependable and Secure Computing (TDSC) *2023*

Complexity Reviewer

Complexity Journal (Complexity) *2022*

SFI Reviewer

Springer Open Financial Innovation (SFI) *2021*

IJAMCS Reviewer

International Journal of Applied Mathematics and Computer Science (IJAMCS) *2021*

Honors & Awards

Summer 2024 **Over 200 Citations in Peer-Reviewed Publications** University of California, Irvine
Fall 2019 **Recipient of $2,500 Discretionary Grant from UC Irvine School of Information and Computer Sciences,** UC Irvine
Summer 2014 **Ranked 22nd Nationally in M.Sc. Entrance Exam among over 30,000 candidates** Tehran, Iran
Summer 2007 **Semifinalist in National Mathematics Olympiad** Tehran, Iran

Coursework

- Machine Learning (A)
- Statistical NLP (A)
- Deep Generative Models (A)
- Algorithms (A)
- Data Structures (A)
- Fundamentals of Computer and Programming (C++) (20/20)
- Advanced Programming (Java) (20/20)
- Algorithms Design (20/20)
- Principles of Database Design (19.75/20)
- Speech Processing & Speech Recognition (A)