

Ali Derakhshan

RESEARCH SCIENTIST · MULTIMODAL LEARNING, TRUSTWORTHY AI & EFFICIENT FOUNDATION MODELS

Irvine, California, United States · U.S. Work Authorized

✉ aderakh1@uci.edu | 📧 aliderakhsh | 🌐 aliderakhsh | 🎓 Google Scholar

Research Summary

I pursue a central question: how can foundation models meet practical constraints without compromising capability or trust? I treat efficiency, safety, and multimodal learning as connected system-design problems rather than isolated benchmarks. My research spans the full cycle: translating concrete challenges into precise learning objectives, designing and implementing methods, and building rigorous evaluation pipelines to test them. This approach has resulted in a co-first-authored paper in ACL Findings 2026, together with peer-reviewed work on robust AI systems and learning across language and vision. With a Ph.D. in Computer Science from UC Irvine and 360+ citations, I bring a demonstrated record of turning research ideas into working, carefully evaluated systems.

Education

University of California, Irvine

PH.D. IN COMPUTER SCIENCE

Irvine, USA

Sep. 2019 – Jun. 2026

- Research Focus: Trustworthy and efficient foundation models, including LLM safety, low-bit quantization, and multimodal learning.

University of California, Irvine

M.SC. IN COMPUTER SCIENCE, GPA: 3.98/4.0

Irvine, USA

Mar 2023

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

Sharif University of Technology

M.SC. IN COMPUTER ENGINEERING, GPA: 3.9/4.0

Tehran, Iran

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

K.N. Toosi University of Technology

B.SC. IN COMPUTER ENGINEERING - HARDWARE,

Tehran, Iran

- Thesis: "Text Summarization Using LSA and NMF"

Technical Skills

Modeling

PyTorch, Hugging Face Transformers, Quanto, HQQ, GPTQ, mixed-precision post-training quantization, adversarial training

Multimodal

CLIP, image-text representation learning, multimodal classification and evaluation

Programming & Data

Python, SQL, NumPy, SciPy, pandas, scikit-learn, TensorFlow, MATLAB, C/C++, Java

Experience

Research Assistant

Prof. Ian Harris

SECURE SYSTEMS AND SOFTWARE LABORATORY, UNIVERSITY OF CALIFORNIA, IRVINE

Sep. 2019 – Jun. 2026

- **CoopQ — Interaction-Aware LLM Quantization (ACL Findings 2026; Co-First Author):** Co-designed an end-to-end, memory-constrained 2/4-bit post-training quantization framework. Designed and implemented SPQE, a sampled-Shapley pipeline that avoids unstable pruning-based estimates through progressive 4-to-2-bit quantization, and defined its per-token negative-log-likelihood (log-perplexity) payoff to estimate layer sensitivities and cross-layer interactions. Evaluated 2B–9B Llama-3, Gemma-2, and Qwen3 models across Quanto, HQQ, and GPTQ, reducing perplexity by 20–80% versus the strongest baselines at matched memory budgets; presented the work at ACL 2026.
- **Adversarial Prompt Shield (WOAH 2024):** Co-developed APS, a lightweight 66M-parameter DistilBERT safety classifier, and BAND, efficiently generated adversarial-training datasets for robustness to unseen gradient-optimized jailbreak suffixes. Evaluated the defense across Vicuna-7B, Falcon-7B, and Guanaco-7B, reducing attack success rate by up to 44.9%; released the code and datasets publicly.
- **CLIP-Based Multimodal Learning (SemEval 2022):** Co-developed a CLIP-based pipeline that encodes and fuses image-text representations for multilabel misogyny identification on the MaMi benchmark, achieving F1 scores of 0.694 and 0.681 on Tasks A and B.
- **Social-Engineering Attack Detection:** Developed semantic-language models using semantic content and speech acts to detect telephone-based social-engineering attacks.
- **Human Study of Telephone Scams:** Designed and conducted a 186-participant study of live telephone scams, created a recorded-call dataset, and analyzed factors associated with susceptibility; published at STAST 2020.
- **NSF-Supported Research:** Developed semantic-language modeling methods and datasets for social-engineering attack detection under NSF Award 1813858, "Detecting Social Engineering Attacks Using Semantic Language Analysis" (PI: Ian Harris).

Additional Research & Teaching Experience

Research Assistant	<i>Prof. Hamid Beigy</i>
SHARIF UNIVERSITY OF TECHNOLOGY, TEHRAN, IRAN	2017 - 2019
• Opinion Mining and Market Prediction: Built machine-learning and probabilistic-inference pipelines for social-media sentiment and stock-movement prediction; published the work in <i>Engineering Applications of Artificial Intelligence</i>. • Master's Thesis: Analyzed purchase satisfaction using opinion mining under the supervision of Prof. Hamid Beigy.	
Teaching Assistant	
DEPARTMENT OF COMPUTER SCIENCE, UNIVERSITY OF CALIFORNIA, IRVINE	2020 – 2024
• Supported graduate courses in Machine Learning, Neural Networks and Deep Learning, and Embedded Systems, plus undergraduate programming and embedded-software courses.	

Publications

Junchen Zhao*, Ali Derakhshan* , Jayden Hyman, Junhao Dong, Sangeetha Abdu Jyothi, Ian Harris	
*Equal contribution	
[1] COOPQ: COOPERATIVE GAME INSPIRED LAYERWISE MIXED PRECISION QUANTIZATION FOR LLMs	2026 - Published
✓ Findings of the Association for Computational Linguistics: ACL 2026, San Diego, CA, USA	
Jinhwa Kim, Ali Derakhshan , Ian Harris	
[2] 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	
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 & Research Impact

Jul. 2026	360+ Google Scholar Citations 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