

Pedram Zaree

pedram.zaree@email.ucr.edu | +1 (951) 462-3503 | [Website](#) | [LinkedIn](#) | [Google Scholar](#)

Education

Ph.D. in Computer Science - Expected 06/2028

09/2022 - Present

University of California, Riverside (UCR), Riverside, CA

GPA: 3.84/4.0

- Advisor: Prof. [Nael Abu-Ghazaleh](#); M.S. in Computer Science earned en route in 06/2024.
- Relevant coursework: Natural Language Processing, Artificial Intelligence, ML, Data Mining, Pattern Recognition.

B.Sc. in Electrical Engineering

09/2017 - 02/2022

Sharif University of Technology, Tehran, Iran

GPA: 3.90/4.0

Research Experience

AI Research Intern - Brave Software

06/2026 - 09/2026

- Built **query-guided visual minimization** for privacy-preserving VQA, improving privacy AP from **0.41 to 0.80** on DIPA2/VISPR while preserving downstream VLM utility.

PhD Research Assistant - University of California, Riverside

09/2022 - Present

- Developing **ModSafe**, which transfers text-trained safety probes to image and audio models via Procrustes alignment; evaluated on Qwen2-Audio for speech safety and benign utility.
- Developed **Attention Eclipse**, an attention-manipulation framework for evaluating LLM safety alignment.
- Developed **AttenMIA**, an attention-based membership-inference framework using head-level signals.
- Developed an attention-based backdoor-detection method for large reasoning models; created and analyzed model variants using supervised fine-tuning, DPO, PPO, LoRA, and PEFT, with model instrumentation and activation extraction.

Publications

Attention Eclipse: Manipulating Attention to Bypass LLM Safety-Alignment

EMNLP 2025

First author; peer-reviewed. Identified an attention-level vulnerability in LLM safety alignment; achieved 91.2% attack success on LLaMA-2 while reducing optimization iterations by 66% relative to prior work.

That Doesn't Go There: Attacks on Shared State in Multi-User AR Applications

USENIX Security '24

Peer-reviewed. Evaluated security failures in interactive multi-user augmented-reality systems and demonstrated end-to-end attacks against shared spatial state in commercial platforms including Google ARCore and Meta Mapillary.

AttenMIA: LLM Membership Inference Attack through Attention Signals

Under Review

First author. First attention-based membership-inference framework; 0.99 ROC AUC and 88% TPR@1% FPR on WikiMIA-32 with LLaMA-2-13B.

Vulnerabilities of Large Language Models to Adversarial Attacks

ACL 2024 Tutorial

A taxonomy of adversarial attacks on LLMs, including jailbreaks, prompt injection, and multimodal threats; presented to 200+ attendees.

Loss Landscape Poisoning: Targeted Extraction of Unseen Training Data from LLMs

arXiv Preprint

Loss-landscape poisoning to extract unseen training secrets, up to **100%** extraction on LLMs and **90%** on VLMs while preserving model utility.

Co(ve)rtex: ML Models as Storage Channels and Their (Mis-)Applications

arXiv Preprint

Over-parameterized ML models can serve as high-capacity covert storage channels without materially affecting primary-task accuracy.

Technical Skills

Programming and ML Frameworks: Python, C/C++; PyTorch, TensorFlow, Hugging Face, scikit-learn, NumPy, pandas, OpenCV.

ML Research: LLMs, multimodal models, computer vision, representation learning, model interpretability, privacy-preserving ML, and adversarial robustness.

Model Adaptation and Efficiency: SFT, DPO, PPO, LoRA, PEFT, RAG, quantization, pruning, knowledge distillation, batching, caching, and memory-constrained inference.

Systems and Evaluation: CUDA, multi-GPU training/inference, vLLM, Linux, Docker, AWS/GCP; latency, throughput profiling.

Additional Experience

Teaching Assistant - UCR and Sharif University of Technology

01/2020 - 04/2026

- Supported Machine Learning, Probability and Statistics, and Introduction to Computing courses.

AI Research Intern - Sharif University of Technology

10/2020 - 08/2021

- CycleGAN pipeline in PyTorch for resolution improvement and noise removal, improving image quality by **35%**.

Selected Honors

2022: Dean's Distinguished Fellowship Award, University of California, Riverside.

2018: Member, Iran's National Elites Foundation.

2017: Full Scholarship Recipient, Sharif University of Technology; ranked 76th nationwide in the university entrance examination among 200,000+ participants.