# Hadi Reisizadeh

+1-612-245-0654 hadi.reisizadeh@gmail.com Website Google Scholar LinkedIn

## Summary

Research Scientist specializing in **LLM safety, unlearning, and optimization**, with first-author work at **ICML** and **EACL** and collaborations with **Amazon AGI**, **Meta**, and **IBM Research**. Develop methods making model safeguards hold up in deployment: exposed knowledge leakage under probabilistic decoding and tool-use recovery through **agent harnesses**, and built training methods (**RULE, BLUR, HARU**) closing these gaps. Results include near-zero leakage under sampling on TOFU (RULE), up to **19**% higher unlearning efficiency (BLUR), harness recovery cut to **4**.**5**% vs. **32**.**8**% (HARU), and **34**% faster reasoning training without accuracy loss (EPiC).

## Technical Skills

| | |
|---|---|
| **Core Expertise** | LLMs, optimization, machine unlearning, efficient training, and safe AI |
| **Programming Languages** | Python, SQL, MATLAB, R, C/C++ |
| **Machine Learning** | PyTorch, HuggingFace Transformers, TensorFlow, Scikit-learn |
| **LLM Training** | DeepSpeed, HuggingFace Accelerate |

## Work Experience

**University of Minnesota**, Minneapolis, MN
Postdoctoral Associate, Optimization for AI Lab — 09/2024 – Present
Research Assistant, Information Processing Group — 09/2017 – 09/2024

- **LLM Unlearning and Model Safety.**
  - Explored the fundamental limits of unlearning in LLMs, showing that sensitive knowledge can resurface under probabilistic decoding. Introduced **leak@**$k$, a meta-metric to quantify worst-case knowledge leakage across $k$ generations, and **RULE**, an iterative training framework achieving near-zero leakage on TOFU and outperforming SOTA methods on MUSE-News [paper], [code].
  - Proposed a **bi-level optimization framework (BLUR)** for LLM unlearning that prioritizes the forget objective over the retain loss, achieving up to **19**% higher unlearning efficiency and **11**.**7**% higher model utility on MUSE-News with the **LLaMA-2-7B** model, and **16**.**6**% higher efficiency on WMDP with equivalent utility using the **Zephyr-7B-Beta** model [paper], [code].
  - Showed that ordinary **agent harnesses** (reasoning guidance, schema repair) recover up to **80**% of tool-use capabilities suppressed by unlearning, without changing model weights. Introduced the **Harness Recovery Rate (HRR)** metric and **HARU**, a harness-aware unlearning method that cuts HRR from **80**% to **0**% on BFCL and from **32**.**8**% to **4**.**5**% on ToolAlpaca while preserving **69**% retained-tool accuracy.
- **Efficient Reasoning Condensation.** Proposed an **edge-preserving condensation framework** for chain-of-thought reasoning training that retains only the initial and final reasoning stages, enabling supervised models to achieve **lossless reasoning accuracy** while reducing training time by over **34**% across multiple **LLM families (Qwen, LLaMA)** and reasoning benchmarks [paper], [code].
- **Distributed Training.** Designed a **two-time-scale optimization algorithm** for communication-efficient distributed training over noisy, time-varying networks, achieving **provable convergence guarantees** in both convex and non-convex settings [paper 1], [paper 2], [paper 3].

**Seagate Technology**, Shakopee, MN — 05/2023 – 04/2024
Research Intern, Data Trust Research Team

- **Differentially Private Distributed Training.** Developed an **accelerated privacy-preserving distributed training method**, achieving faster convergence without reducing privacy guarantees.

## Education

**Ph.D. & M.Sc. in Electrical and Computer Engineering** (*Minor in Computer Science*)
*University of Minnesota, Twin Cities*, Minneapolis, MN — 09/2017 – 09/2024

## Selected Publications

- **H. Reisizadeh**, J. Ruan, Y. Chen, S. Pal, S. Liu, and M. Hong, *"Leak@*$k$*: Unlearning Does Not Make LLMs Forget Under Probabilistic Decoding,"* ICML, 2026.
- **H. Reisizadeh**, J. Jia, Z. Bu, B. Vinzamuri, A. Ramakrishna, K. Chang, V. Cevher, S. Liu, and M. Hong, *"BLUR: A Bi-Level Optimization Approach for LLM Unlearning,"* EACL, 2026.
- **H. Reisizadeh**, J. Ruan, S. Liu, and M. Hong, *"Do LLMs Really Forget? Hidden-State Leakage in Model Unlearning and How to Fix it,"* Submitted, 2026.
- J. Ruan, **H. Reisizadeh**, S. Liu, and M. Hong, *"When Unlearning Meets the Harness: Helpful Assistance, Harmful Recovery,"* Submitted, 2026.
- J. Jia, **H. Reisizadeh**, C. Fan, N. Baracaldo, M. Hong, and S. Liu, *"EPiC: Towards Lossless Speedup for Reasoning Training through Edge-Preserving CoT Condensation,"* Submitted, 2026.