Abhinav Kumar

securegradients.com | abhinavk@umass.edu

Research Interests

My research focuses on developing trustworthy AI systems for real-world applications, with a particular emphasis on the security and privacy of generative reasoning models in web-agent and multi-agent settings.

Education

University of Massachusetts Amherst Sept 2024 – May 2028
Ph.D. in Computer Science; Advisor: Eugene Bagdasarian

Saint Louis University Aug 2023 – Jul 2024
M.S. in Artificial Intelligence; Advisor: Reza Tourani

Saint Louis University Aug 2019 – Jul 2023
B.S. in Computer Engineering

Experience

Research Intern, Max Planck Institute for Intelligent Systems June 2026 – Present

- Researching safety risks of miscoordination in multi-agent systems.

Graduate Research Assistant, University of Massachusetts Amherst Sept 2024 – Present

- Researching trustworthy AI systems by studying potential exploits and defenses.
- Creating frameworks to benchmark and study agent and multi-agent safety.

External Collaborator, ServiceNow Feb 2026 – Present

- Creating benchmarks for multi-agent contextual integrity.

Graduate Research Assistant, Saint Louis University Aug 2023 – Jul 2024

- Conducted research on privacy-preserving machine learning, MLaaS validation, and trustworthy AI systems.

External Collaborator, Intel Labs August 2021 – Jul 2024

- Developed frameworks and algorithms for pervasive edge privacy.

Undergraduate Research Assistant, Saint Louis University May 2021 – Jul 2023

- Developed trusted hardware solutions for machine learning at the edge.

Awards & Grants

- IEEE/IFIP DSN Best Paper Award (Aug 2025)
- IEEE PerCom Best Industry Track Paper Award (Mar 2022)
- Saint Louis University Scholarship (Aug 2019)

Publications

Preprints

Colosseum: Auditing Collusion in Cooperative Multi-Agent Systems.

Mason Nakamura*, **Abhinav Kumar***, Saswat Das*, Sahar Abdelnabi, Saaduddin Mahmud, Ferdinando Fioretto, Shlomo Zilberstein, Eugene Bagdasarian.

arXiv, 2026. (*Equal Contribution)

Terrarium: Revisiting the Blackboard for Multi-Agent Safety, Privacy, and Security Studies.

Mason Nakamura*, **Abhinav Kumar***, Saaduddin Mahmud, Sahar Abdelnabi, Shlomo Zilberstein, Eugene Bagdasarian.

arXiv, 2025. (*Equal Contribution)

Throttling Web Agents Using Reasoning Gates.

Abhinav Kumar, Jaechul Roh, Ali Naseh, Amir Houmansadr, Eugene Bagdasarian.

arXiv, 2025.

OverThink: Slowdown Attacks on Reasoning LLMs.

Abhinav Kumar, Jaechul Roh, Ali Naseh, Marzena Karpinska, Mohit Iyyer, Amir Houmansadr, Eugene Bagdasarian.

arXiv, 2025.

Peer-Reviewed Conferences

Network-Level Prompt and Trait Leakage in Local Research Agents.

Hyejun Jeong, Mohammadreza Teymoorianfard, **Abhinav Kumar**, Amir Houmansadr, Eugene Bagdasarian.

USENIX Security 2026, ICML AIWILD Workshop 2026.

Privacy Analysis of Oblivious DNS over HTTPS: a Website Fingerprinting Study.

Mohammad Salari, **Abhinav Kumar**, Federico Rinaudi, Reza Tourani, Alessio Sacco, Flavio Esposito.

IEEE/IFIP DSN 2025. **Best Paper Award**.

Persistent Backdoor Attacks in Continual Learning.

Zhen Guo, **Abhinav Kumar**, Reza Tourani.

USENIX Security 2025.

Silver Linings in the Shadows: Harnessing Membership Fingerprinting for Machine Unlearning.

Nexhi Sula, **Abhinav Kumar**, Jie Hou, Han Wang, Reza Tourani.

IEEE CNS 2025.

A Generative Framework for Low-Cost Result Validation of Machine Learning-as-a-Service Inference.

Abhinav Kumar, Miguel A. Guirao Aguilera, Reza Tourani, Satyajayant Misra.

ACM ASIACCS 2024.

A Stealthy Inference Attack on Split Learning with a Split-Fuse Defensive Measure.

Sean Dougherty, **Abhinav Kumar**, Jie Hou, Reza Tourani, Atena M. Tabakhi.

IEEE CNS 2023.

SCLERA: A framework for Privacy Preserving MLaaS at the Pervasive Edge.

Abhinav Kumar, Reza Tourani, Mona Vij, Srikathyayani Srikanteswara.

IEEE PerCom 2022. **Best Industry Track Paper Award**.

Invited Talks

OverThink: Slowdown Attacks on Reasoning LLMs.

Google Adversarial ML Knowledge Sharing, Mar 2025.

DcTC: Towards Data-centric Trusted Computing.

Intel SPR Tech Talks, Jul 2025.

A Generative Framework for Result Validation of Outsourced Machine Learning Workloads via TEE.

NSF RINGS Workshop, Jun 2023.