

Sahar Abdelnabi

AI/ML Security and Safety Researcher

✉ sahar.s.abdelnabi@gmail.com

🏠 <https://s-abdelnabi.github.io/>

🔗 [Google Scholar](#)

RESEARCH FOCUS

I am interested in the broad intersection of machine learning with security, safety, and sociopolitical aspects. This includes the following areas: 1) Understanding, probing, and mitigating the failure modes of machine learning models, their biases, and their misuse scenarios. 2) How machine learning models could amplify or help counter existing societal and safety problems (e.g., misinformation, biases, and stereotypes). 3) Emergent safety challenges posed by new foundation and large language models.

CURRENT POSITION

AI Security Researcher

2024 - Ongoing

Microsoft Security Response Center (MSRC), Microsoft Research Cambridge, UK

Role:

- Conducting AI security and safety research
- Assessing AI security vulnerabilities reported through Microsoft's AI Bug Bounty program

EDUCATION

Ph.D. in Computer Science

2019 - 2024

CISPA Helmholtz Center for Information Security, Germany

Advisor: Prof. Dr. Mario Fritz

Grade: **Summa cum laude** (External reviewers: Battista Biggio and Florian Tramèr)

M.Sc. in Computer Science

2017 - 2019

Saarland University, Germany

GPA: 1.2/1.0 (Thesis GPA: 1.0)

Advisor: Prof. Dr. Mario Fritz

M.Sc. in Computer and System Engineering

2013 - 2017

Ain Shams University, Egypt

GPA: 3.7/4.0

B.Sc. in Computer and System Engineering

2008 - 2013

Ain Shams University, Egypt

Miscellaneous

[CISPA Summer School on Trustworthy AI](#), 2022

[ELLIS Doctoral Symposium on "AI for Good"](#), 2022

PREVIOUS RESEARCH AND INDUSTRY EXPERIENCE

PhD Candidate

2019 - 2024

CISPA Helmholtz Center for Information Security, Germany

Research Assistant

2017 - 2019

Max Planck Institute for Informatics, Germany

Advised by Prof. Dr. Andreas Bulling

Quality Assurance Engineer

2013 - 2017

Mentor Graphics (Currently, Siemens EDA), Egypt

TEACHING EXPERIENCE

Teaching Assistant

2020- 2023

Saarland University, Germany

Classes: “Opportunities and Risks of Large Language Models and Foundation Models” seminar,
“Machine Learning in Cybersecurity”, “High-level Computer Vision”

Tutor

2017- 2019

Saarland University, Germany

Classes: “Image Processing and Computer Vision”, “Neural Networks Implementation and Applications”,
“Interactive Systems”

PUBLICATIONS

- [1] **Sahar Abdelnabi***, Amr Gomaa*, Eugene Bagdasarian, Per Ola Kristensson, and Reza Shokri. “Firewalls to Secure Dynamic LLM Agentic Networks”. In: *Arxiv*. 2025.
- [2] Ivaxi Sheth, Jan Wehner*, **Sahar Abdelnabi***, Ruta Binkyte*, and Mario Fritz. “Safety is Essential for Responsible Open-Ended Systems”. In: *Arxiv*. 2025.
- [3] **Sahar Abdelnabi**, Amr Gomaa, Sarath Sivaprasad, Lea Schönherr, and Mario Fritz. “Cooperation, Competition, and Maliciousness: LLM-Stakeholders Interactive Negotiation”. In: *NeurIPS Datasets and Benchmarks*. 2024.
- [4] Edoardo Debenedetti, Javier Rando, Daniel Paleka, Silaghi Fineas Florin, Dragos Albastroiu, Niv Cohen, Yuval Lemberg, Reshmi Ghosh, Rui Wen, Ahmed Salem, Giovanni Cherubin, Santiago Zanella-Beguelin, Robin Schmid, Victor Klemm, Takahiro Miki, Chenhao Li, Stefan Kraft, Mario Fritz, Florian Tramèr, **Sahar Abdelnabi**, and Lea Schönherr. “Dataset and Lessons Learned from the 2024 SaTML LLM Capture-the-Flag Competition”. In: *NeurIPS Datasets and Benchmarks. Spotlight*. 2024.
- [5] **Sahar Abdelnabi***, Aideen Fay*, Giovanni Cherubin, Ahmed Salem, Mario Fritz, and Andrew Paverd. “Are you still on track!? Catching LLM Task Drift with Activations”. In: *SaTML*. To appear. 2025.
- [6] Egor Zverev, **Sahar Abdelnabi**, Mario Fritz, and Christoph H Lampert. “Can LLMs Separate Instructions From Data? And What Do We Even Mean By That?”. In: *ICLR*. 2025.
- [7] Ivaxi Sheth, **Sahar Abdelnabi**, and Mario Fritz. “Hypothesizing Missing Causal Variables with LLMs”. In: *Causality and Large Models NeurIPS Workshop*. 2024.
- [8] Sarath Sivaprasad, Pramod Kaushik, **Sahar Abdelnabi**, and Mario Fritz. “Exploring Value Biases: How LLMs Deviate Towards the Ideal”. In: *LLMs and Cognition ICML Workshop*. 2024.
- [9] **Sahar Abdelnabi***, Kai Greshake*, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. “Not what youve signed up for: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection”. In: *AISec Workshop, in conjunction with CCS*. *: Equal contribution. **Oral Presentation, Best Paper Award**. 2023.
- [10] **Sahar Abdelnabi** and Mario Fritz. “Fact-Saboteurs: A Taxonomy of Evidence Manipulation Attacks against Fact-Verification Systems”. In: *USENIX Security*. 2023.
- [11] Giada Stivala, **Sahar Abdelnabi**, Andrea Mengascini, Mariano Graziano, Mario Fritz, and Giancarlo Pellegrino. “From Attachments to SEO: Click Here to Learn More about Clickbait PDFs!”. In: *Annual Computer Security Applications Conference (ACSAC)*. 2023.
- [12] Rebecca Weil, **Sahar Abdelnabi**, Mario Fritz, and Rakibul Hasan. “Tell me what you like and I know what you will share: Topical interest influences behavior toward news from high and low credible sources”. In: *EuroS&P Workshops*. 2024.
- [13] **Sahar Abdelnabi**, Rakibul Hasan, and Mario Fritz. “Open-Domain, Content-based, Multi-modal Fact-checking of Out-of-Context Images via Online Resources”. In: *CVPR*. 2022.
- [14] **Sahar Abdelnabi** and Mario Fritz. “Adversarial Watermarking Transformer: Towards Tracing Text Provenance with Data Hiding”. In: *S&P*. 2021.
- [15] Ning Yu, Vladislav Skripniuk, **Sahar Abdelnabi**, and Mario Fritz. “Artificial fingerprinting for generative models: Rooting deepfake attribution in training data”. In: *ICCV. Oral presentation*. 2021.

- [16] **Sahar Abdelnabi** and Mario Fritz. “What’s in the box: Deflecting Adversarial Attacks by Randomly Deploying Adversarially-Disjoint Models”. In: *the 8th ACM Workshop on Moving Target Defense, in conjunction with CCS*. 2021.
- [17] **Sahar Abdelnabi**, Katharina Krombholz, and Mario Fritz. “Visualphishnet: Zero-day phishing website detection by visual similarity”. In: *CCS*. 2020.
- [18] **Sahar Abdelnabi**, Michael Xuelin Huang, and Andreas Bulling. “Towards High-Frequency SSVEP-Based Target Discrimination with an Extended Alphanumeric Keyboard”. In: *IEEE International Conference on Systems, Man and Cybernetics (SMC)*. 2019.
- [19] **Sahar Abdelnabi**, Seif Eldawlatly, and Mahmoud I Khalil. “Epileptic seizure prediction using zero-crossings analysis of EEG wavelet detail coefficients”. In: *IEEE conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*. 2016.

ACADEMIC ACTIVITIES

Competition Organizations

One of the main organizers of IEEE SaTML’25 challenge “LLMail-Inject: Adaptive Prompt Injection Attacks”
 Co-organizer of the IEEE SaTML’24 challenge “LLM CtF”

Reviewing and Consulting

PC member of IEEE SaTML’24, AISec’23 and’24 Workshop, USENIX Security’25, AAAI’25, CCS’25
 Reviewer for ICLR’25, ICML’24, ICLR’24, NeurIPS’23, ICML’23 Neural Conversational AI Workshop, ICCV’23, CVPR’23, ECCV’22, CVPR’22, IEEE TPAMI (2024, 2022, 2021), ICLR’21 Workshop on Synthetic Data Generation Quality, Privacy, Bias
 Grant reviewing for [Cooperative AI](#)
 Consulting for the [UK AI Safety Institute](#)

Talks

Towards Aligned, Interpretable, and Steerable Safe AI Agents
 TU Graz 2025, and UMass Amherst Security and Privacy Seminar 2025
On the Security of Real-World LLM-Integrated Applications
 Invited talk at the [European Symposium on Security and Artificial Intelligence](#), 2024
On New Security and Safety Challenges Posed by LLMs and How to Evaluate Them
 Keynote at [HIDA PhD Meet-up](#) and a talk at [MLSec seminars](#), 2024
On Evaluating Language Models and Their Security and Safety Implications
 Vector Institute and ETH Zürich, 2023
LLM-Deliberation: Evaluating LLMs with Interactive Multi-Agent Negotiation Games
[SIGSEC talk](#), 2023
Compromising LLMs: The Advent of AI Malware
 Black Hat USA, 2023
Panel: Security of Generative AI and Generative AI in Security
 Invited panelists at DIMVA, 2023
Not what you’ve signed up for: Investigating the Security of LLM-Integrated Applications
 Privacy and Security in ML Seminars, 2023
How to Improve Automated Fact-Checking
 Max Planck Institute for Software Systems, 2022
Multi-modal Fact-checking: Out-of-Context Images and How to Catch Them
 UCL Information Security seminars, 2022

AWARDS

Best paper award and oral presentation at AISec, 2023
 Academic Research Grant for Google Cloud research credits, 2023

Academic Research Grant for Google Cloud research credits, 2021

Oral presentation at ICCV, 2021

Saarland University scholarship for international students (DAAD STIBET III scholarship grant), 2019

IEEE Computational Intelligence Society (CIS) Outstanding Student-Paper Travel Grant for CIBCB, 2016

OUTREACH

Podcasts, documentaries, and community events

Research Cohort - Implementation and Evaluation of a Research Paper

Invited panelist

AI Saturdays Lagos, 2024

ChatGPT: What happens when the AI takes over?

Y-Kollektiv Documentary, 2023

A dark side to LLMs

CyberWire Podcast, 2023

Deepfakes and Fingerprinting

CISPA tl;dr Podcast, 2022

Interviews

Our work on “indirect prompt injection” has been featured as interviews with myself/authors in [Vice](#), [Wired](#), [Zeit](#), [MIT Technology Review](#), and [CISPA](#) communication channels.

Our work on “LLM-deliberation” has been featured by [the Montreal AI Ethics Institute](#).

Other press and community coverage

Our work on “indirect prompt injection” has been featured by policymakers and practitioners such as the [German Federal Office for Information Security](#), [NIST](#), [OWASP](#), [MITRE](#), [Microsoft's AI bug bar](#), and many others, in addition to introducing new terminologies for the entire research and tech fields.

LANGUAGES

English (Bilingual)

German (Intermediate)

Colloquial Egyptian (Mother Tongue)