Sahar Abdelnabi

AI/ML Security and Safety Researcher

✉ sahar.abdelnabi@tue.ellis.eu
https://s-abdelnabi.github.io/
Google Scholar

Research Focus

I am interested in the broad intersection of AI, AGI, multi-agent systems with security, safety, and sociopolitical aspects. This includes the following areas:

1) Understanding, probing, and evaluating the failure modes of AI models, their biases, emergent risks, and their misuse scenarios.
2) How to design mitigations, system defenses, white-box control methods, reasoning enhancements, and novel architectures to counter such risks.
3) Leveraging AI agents for good: scientific discovery and advancing our society.

Our previous work, in 2023, was the first to identify the indirect prompt injection vulnerability in LLM-integrated applications, and in 2020, to propose and call for watermarking generative AI. Our work on LLM sampling heuristics received a Best Paper Award at ACL2025.

Current Position

Principal Investigator and an Independent Research Group Leader *Oct. 2025 -*
ELLIS Institute Tübingen, the Max-Planck Institute for Intelligent Systems, and Tübingen AI Center, Germany
Role:
- I lead the COMPASS (COoperative Machine intelligence for People-Aligned Safe Systems) research group, focused on developing safe, aligned, and steerable AI agents with emphasis on security, human aspects, and cooperative multi-agent systems.
- I am a core faculty member of The International Max Planck Research School for Intelligent Systems (IMPRS-IS) and an associated faculty member of the Max Planck ETH Center for Learning Systems (CLS).

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: Prof. Dr. Battista Biggio and Prof. Dr. 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

Previous Research and Industry Experience

AI Security Researcher (Full-Time Employee) *Feb. 2024 - Sept. 2025*
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

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] Mason Nakamura*, Abhinav Kumar*, Saaduddin Mahmud, **Sahar Abdelnabi**, Shlomo Zilberstein, and Eugene Bagdasarian. “Terrarium: Revisiting the Blackboard for Multi-Agent Safety, Privacy, and Security Studies”. In: *arXiv* (2025).

[2] **Sahar Abdelnabi**, Aideen Fay, Ahmed Salem, Egor Zverev, Kai-Chieh Liao, Chi-Huang Liu, Chun-Chih Kuo, Jannis Weigend, Danyael Manlangit, Alex Apostolov, et al. “LLMail-Inject: A Dataset from a Realistic Adaptive Prompt Injection Challenge”. In: *arXiv* (2025).

[3] **Sahar Abdelnabi** and Ahmed Salem. “The Hawthorne Effect in Reasoning Models: Evaluating and Steering Test Awareness”. In: ***NeurIPS 2025*** *(To appear)*. **Spotlight**. 2025.

[4] Guangchen Lan, Huseyin A. Inan, **Sahar Abdelnabi**, Janardhan Kulkarni, Lukas Wutschitz, Reza Shokri, Christopher G. Brinton, and Robert Sim. “Contextual Integrity in LLMs via Reasoning and Reinforcement Learning”. In: ***NeurIPS 2025*** *(To appear)*. 2025.

[5] Jan Wehner, **Sahar Abdelnabi**, Daniel Tan, David Krueger, and Mario Fritz. “Taxonomy, Opportunities, and Challenges of Representation Engineering for Large Language Models”. In: ***TMLR*** (2025). **Survey Certification**.

[6] **Sahar Abdelnabi***, Amr Gomaa*, Eugene Bagdasarian, Per Ola Kristensson, and Reza Shokri. “Firewalls to Secure Dynamic LLM Agentic Networks”. In: *Arxiv*. 2025.

[7] Ivaxi Sheth, **Sahar Abdelnabi**, and Mario Fritz. “Context-Aware Reasoning On Parametric Knowledge for Inferring Causal Variables”. In: ***EMNLP findings***. 2025.

[8] Ivaxi Sheth, Jan Wehner*, **Sahar Abdelnabi***, Ruta Binkyte*, and Mario Fritz. “Safety is Essential for Responsible Open-Ended Systems”. In: *ICLR Workshops*. 2025.

[9] Sarath Sivaprasad*, Pramod Kaushik*, **Sahar Abdelnabi**, and Mario Fritz. “A Theory of Response Sampling in LLMs: Part Descriptive and Part Prescriptive”. In: ***ACL (main)***. **Oral & Panel Discussion & A Best Paper Award**. 2025.

[10] **Sahar Abdelnabi***, Aideen Fay*, Giovanni Cherubin, Ahmed Salem, Mario Fritz, and Andrew Paverd. “Get My Drift? Catching LLM Task Drift with Activation Deltas”. In: ***SaTML***. 2025.

[11] 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.

[12] **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.

[13] 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.

[14] Kai Greshake*, **Sahar Abdelnabi***, 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.

[15] **Sahar Abdelnabi** and Mario Fritz. “Fact-Saboteurs: A Taxonomy of Evidence Manipulation Attacks against Fact-Verification Systems”. In: ***USENIX Security***. 2023.

[16] 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.

[17] 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.

[18] **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.

[19] **Sahar Abdelnabi** and Mario Fritz. "Adversarial Watermarking Transformer: Towards Tracing Text Provenance with Data Hiding". In: ***S&P***. 2021.

[20] 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.

[21] **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.

[22] **Sahar Abdelnabi**, Katharina Krombholz, and Mario Fritz. "Visualphishnet: Zero-day phishing website detection by visual similarity". In: ***CCS***. 2020.

[23] **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.

[24] **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 and Services

Workshop Organizations

- Co-organizer of the Foundations of Language Model Security Workshop at EurIPS'25

Competition Organizations

- One of the main and lead 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 SaTML('24-'25), AISec Workshop ('23-'25), USENIX Security'25, AAAI'25, CCS'25, S&P'25

- Reviewer for NeurIPS'25, 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

Mentoring

- I am serving as a mentor in the upcoming (Cooperative AI 2025 research fellowship)

Talks

- *What does it mean for AI agents to preserve privacy?*
—— Graz Security Week (Summer school), 2025

- *Firewalls to Secure Dynamic LLM Agentic Networks*
—— Brave, 2025
—— Google DeepMind, 2025
—— Qualcomm, 2025

- *Panel: Women in AI Security Workshop*
—— The Alan Turing Institute, 2025

- *Towards Aligned, Interpretable, and Steerable Safe AI Agents*
—– TU Graz, 2025
—– UMass Amherst Security and Privacy Seminar, 2025
—– CISPA, 2025
—– ELLIS Institute Tübingen Scientific Symposium, 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, 2024
—– MLSec seminars, 2024

- *On Evaluating Language Models and Their Security and Safety Implications*
—– Vector Institute, 2023
—– 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

- Oral presentation, panel discussion, and a **Best Paper Award** at ACL, 2025 (awarded to 4 out of more than 3000 accepted papers)

- Vector Distinguished Postdoctoral Fellowship, 2023 (declined)

- **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 and blogs

- 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.

- Microsoft Security Response Center (MSRC)’s blog: Announcing the Adaptive Prompt Injection Challenge (LLMail-Inject) and Announcing the winners of the Adaptive Prompt Injection Challenge (LLMail-Inject)

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 (Fluent/Bilingual)
- German (Intermediate)
- Colloquial Egyptian (Mother Tongue)