David Pape

PhD Researcher in AI Security

david.pape@cispa.de | Website | Google Scholar

Research Interests

Security and trustworthiness of large language models and AI agents, LLM evaluation and reproducibility.

Education

Ph.D. in Computer Science (in progress) 2022–present
CISPA Helmholtz Center for Information Security, Saarbrücken, Germany
Advisor: Dr. Lea Schönherr

M.Sc. in IT Security / Information Technology 2019–2022
Ruhr-Universität Bochum, Germany
Final grade: 1.1 (German scale; 1.0 is best)

B.Sc. in IT Security / Information Technology 2015–2019
Ruhr-Universität Bochum, Germany

Publications

The Silent Hyperparameter: Quantifying the Impact of Inference Backends on LLM Reproducibility
David Pape, Jonathan Evertz, and Lea Schönherr
Advances in Neural Information Processing Systems (NeurIPS), 2026. | Paper

No More, No Less: Task Alignment in Terminal Agents
Sina Mavali, **David Pape**, Jonathan Evertz, Samira Abedini, Devansh Srivastav, Thorsten Eisenhofer, Sahar Abdelnabi, and Lea Schönherr
NeurIPS 2026, Evaluations & Datasets Track. | Paper

Unknown Unknowns: Do Hidden Intentions in LLMs Evade Detection?
Devansh Srivastav, **David Pape**, and Lea Schönherr
Findings of the Association for Computational Linguistics: EMNLP 2026 | Paper

Adversarial Robustness of AI-Generated Image Detectors in the Real World
Sina Mavali, Jonas Ricker, **David Pape**, Asja Fischer, and Lea Schönherr
International Conference on Detection of Intrusions and Malware & Vulnerability Assessment (DIMVA), 2026 | Paper

Chasing Shadows: Pitfalls in LLM Security Research
Jonathan Evertz, Niklas Risse, Nicolai Neuer, Andreas Müller, Philipp Normann, Gaetano Sapia, Srishti Gupta, **David Pape**, Soumya Shaw, Devansh Srivastav, Christian Wressnegger, Erwin Quiring, Thorsten Eisenhofer, Daniel Arp, and Lea Schönherr
Network and Distributed System Security Symposium (NDSS), 2026 | Paper

Prompt Obfuscation for Large Language Models
David Pape, Sina Mavali, Thorsten Eisenhofer, and Lea Schönherr
34th USENIX Security Symposium (USENIX Security), 2025 | Paper

On the Limitations of Model Stealing with Uncertainty Quantification Models
David Pape, Sina Däubener, Thorsten Eisenhofer, Antonio Emanuele Cinà, and Lea Schönherr
European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN), 2023 | Paper

Preprints

Contextual Candour: Role-Aware Graduated Disclosure for LLMs over Mixed-Sensitivity Contexts
Devansh Srivastav, **David Pape**, Samira Abedini, and Lea Schönherr
arXiv:2610.28067, 2026. under submission | Paper

Academic Service

Conference reviewing: CYBERSEC 2026; ICML 2026 (**Gold Reviewer**); NeurIPS 2026; AAAI 2027.

Workshop reviewing: AISec (ACM CCS 2024, 2025, 2026; **Top Reviewer, 2025**); DeMeSSAI (IEEE EuroS&P 2025, 2026); FLMSec (NeurIPS 2026).

Artifact evaluation: USENIX Security 2026.

Teaching and Supervision

Teaching Assistant, Trustworthy Machine Learning Summer 2025, 2026
University of Hamburg, Germany

Teaching Assistant, Trustworthy Generative Machine Learning (Seminar) Summer 2024
Saarland University, Germany

Teaching Assistant, Security Courses 2021-2022
Ruhr-Universität Bochum, Germany
Chairs of Systems Security, Embedded Security, and Cryptology & IT Security

Thesis Supervision
Saarland University, Germany
1x Bachelor Thesis, 1x Master Thesis

Talks and Presentations

Protecting, Executing, and Auditing: Security Challenges Across the LLM Lifecycle June 2026
AI Security Afternoon, organized by AI Safety Saarland, Saarbrücken, Germany.

Skills & Languages

Programming & ML	Python, PyTorch, Hugging Face transformers
LLM Systems	LLM inference backends, evaluation and benchmarking
Development	Git, Docker, AI-assisted coding
Languages	German (native), English (C1)