

Nils Lukas

Assistant Professor • MBZUAI • Abu Dhabi, UAE
nils.lukas@mbzuai.ac.ae • nilslukas.github.io • [Gscholar](#)
Updated on November 26, 2025

| | | | |
|-------------------------|--|---|--|
| Research Interests | Design secure and private Machine Learning systems in the presence of untrustworthy | | |
| | <div>1. Providers: Confidential computing via Homomorphic Encryption & Secret Sharing.</div> <div>2. Data: Mitigate data poisoning during training & prompt injection during inference.</div> <div>3. Models: Protect training data privacy through PII scrubbing & differential privacy.</div> <div>4. Users: Control misuse by detecting generated (mis)information with watermarking.</div> | | |
| Education | University of Waterloo, Canada | 2019 - 02/2024 | |
| | Ph.D. in Computer Science | | |
| | <div>▪ Advisor: Florian Kerschbaum</div> <div>▪ Thesis: Analyzing Threats of Large-Scale Machine Learning Systems</div> <div>▪ Thesis Awards: Top Mathematics Doctoral Prize & Alumni Gold Medal</div> | | |
| | RWTH-Aachen, Germany | | |
| | M.Sc. in Computer Science (w/Distinction) | 2016 - 2018 | |
| | B.Sc. in Computer Science | 10/2012 - 2016 | |
| Honors & Awards | First Place at the NeurIPS'24 Watermarking Competition [4,400 USD] | 2024 | |
| | First Place at DGE Elite Hackathon, GITEX'24 [10,900 USD] | 2024 | |
| | Top Mathematics Doctoral Thesis, University of Waterloo [1 080 USD] | 2024 | |
| | Alumni Gold Medal, One PhD Award Yearly, University of Waterloo | 2024 | |
| | Best Poster Award, Sponsored by David R. Cheriton [220 USD] | 2023 | |
| | Distinguished Contribution Award, Microsoft MLADS conference | 2023 | |
| | David R. Cheriton Scholarship, University of Waterloo [14 400 USD] | 2022, 2023 | |
| | Outstanding Reviewer, ICML'22 | 2022 | |
| | Best Poster Award, Sponsored by Rogers [720 USD] | 2019 | |
| Conference Publications | [NeurIPS'25] | Mask Image Watermarking | |
| | AR: 24.5% (5 290/21 575) | Runyi Hu, Jie Zhang, Shiqian Zhao, Nils Lukas, Jiwei Li, Qing Guo, Han Qiu, Tianwei Zhang. The Thirty-Ninth Annual Conference on Neural Information Processing Systems, 2025. | |
| | [EMNLP'25] | SPIRIT: Patching Speech Language Models against Jailbreak Attacks | |
| | AR: 22.2% (1 811/8 172) | Amir Djanibekov, Nurdaulet Mukhituly, Kentaro Inui, Hanan Aldarmaki, Nils Lukas. Empirical Methods in Natural Language Processing (Main Conference), 2025. | |
| | [ICML'25] | Optimizing Adaptive Attacks against Content Watermarks for Language Models | |
| | AR: 26.9% (3 260/12 107) | Abdulrahman Diaa, Toluwani Aremu, Nils Lukas. The Forty-Second International Conference on Machine Learning, 2025. | |
| | 🏆 Spotlight (Top 2.6%) | | |
| | [ICML'25] | Cowpox: Towards the Immunity of VLM-based Multi-Agent Systems | |
| | AR: 26.9% (3 260/12 107) | Yutong Wu, Jie Zhang, Yiming Li, Chao Zhang, Qing Guo, Han Qiu, Nils Lukas, Tianwei Zhang. The Forty-Second International Conference on Machine Learning, 2025. | |
| | [USENIX'24] | PEPSI: Practically Efficient Private Set Intersection in the Unbalanced Setting | |
| | AR: 19.1% (417/2 176) | Rasoul Mahdavi, Nils Lukas, Faezeh Ebrahimianghazani, Thomas Humphries, Bailey Kacsmar, John Premkumar, Xinda Li, Simon Oya, Ehsan Amjadian, Florian Kerschbaum. In the 33rd USENIX Security Symposium, 2024. | |

| | | |
|----------------------|--|---|
| Journal Publications | [USENIX'24] AR: 19.1% (417/2176) | Fast and Private Inference of Deep Neural Networks by Co-designing Activation Functions Abdulrahman Diao, Lucas Fenaux, Thomas Humphries, Marian Dietz, Faezeh Ebrahimiaghazani, Bailey Kacsmar, Xinda Li, Nils Lukas , Rasoul Akhavan Mahdavi, Simon Oya, Ehsan Amjadian, Florian Kerschbaum. In the 33rd USENIX Security Symposium, 2024. |
| | [ICLR'24] AR: 30.8% (2250/7262) | Leveraging Optimization for Adaptive Attacks on Image Watermarks Nils Lukas , Abdulrahman Diao, Lucas Fenaux, Florian Kerschbaum. In the Twelfth International Conference on Learning Representations, 2024. |
| | [ICLR'24] AR: 30.8% (2250/7262) 🌐 Media Coverage | Universal Backdoor Attacks Benjamin Schneider, Nils Lukas , Florian Kerschbaum. In the Twelfth International Conference on Learning Representations, 2024. |
| | [USENIX'23] AR: 29.2% (422/1444) | PTW: Pivotal Tuning Watermarking for Pre-Trained Image Generators Nils Lukas and Florian Kerschbaum. In the 32nd USENIX Security Symposium, 2023. |
| | [S&P'23] AR: 17.0% (195/1147) 🏆 Distinguished Contribution Award at Microsoft MLADS | Analyzing Leakage of Personally Identifiable Information in Language Models Nils Lukas , Ahmed Salem, Robert Sim, Shruti Tople, Lukas Wutschitz, Santiago Zanella-Béguelin. In the 44th IEEE Symposium on Security and Privacy, 2023. |
| | [S&P'22] AR: 14.5% (147/1012) | SoK: How Robust is Image Classification Deep Neural Network Watermarking? Nils Lukas , Edward Jiang, Xinda Li, Florian Kerschbaum. In the 43rd IEEE Symposium on Security and Privacy, 2022. |
| | [ICLR'21] AR: 28.7% (860/2997) 🌟 Spotlight (Top 5%) | Deep Neural Network Fingerprinting by Conferrable Adversarial Examples Nils Lukas , Yuxuan Zhang, Florian Kerschbaum. The Ninth International Conference on Learning Representations, 2021. |
| | [IH&MMSEC'21] AR: 40.3% (128/318) | On the Robustness of Backdoor-based Watermarking in Deep Neural Networks Masoumeh Shafieinejad, Nils Lukas , Jiaqi Wang, Xinda Li, Florian Kerschbaum. Proceedings of the 2021 ACM Workshop on Information Hiding and Multimedia Security, 2021. |
| | [ACSAC'20] AR: 20.9% (104/497) | Practical Over-Threshold Multi-Party Private Set Intersection Rasoul Mahdavi, Thomas Humphries, Bailey Kacsmar, Simeon Krastnikov, Nils Lukas , John Premkumar, Masoumeh Shafieinejad, Simon Oya, Florian Kerschbaum, Erik-Oliver Blass. Annual Computer Security Applications Conference (ACSAC), 2020. |
| | [EuroS&P'20] AR: 20.9% (39/187) | Differentially Private Two-Party Set Operations Bailey Kacsmar, Basit Khurram, Nils Lukas , Alexander Norton, Masoumeh Shafieinejad, Zhiwei Shang, Yaser Baseri, Maryam Sepehri, Simon Oya, Florian Kerschbaum. IEEE European Symposium on Security and Privacy (EuroS&P), 2020. |
| | [AIP'18] | SunFlower: A new Solar Tower Simulation Method for use in Field Layout Optimization , Pascal Richter, Gregor Heiming, Nils Lukas , Martin Frank. AIP Conference Proceedings, Volume 2033, Issue 1, 2018. |

| | | |
|-----------------|---|--|
| Workshop Papers | [GenAI4Health] | Sanitizing Medical Documents with Differential Privacy using Large Language Models Rushil Thareja, Gautam Gupta, Preslav Nakov, Praneeth Vepakomma, Nils Lukas . 2025. |
| | [WMARK'25] | First-Place Solution to NeurIPS 2024 Invisible Watermark Removal Challenge Fahad Shamshad, Tameem Bakr, Yahia Salaheldin Shaaban, Noor Hazim Hussein, Karthik Nandakumar and Nils Lukas . The 1st Workshop on GenAI Watermarking, 2025. |
| | [WMARK'25] 🏆 Oral Presentation | Optimizing Adaptive Attacks against Content Watermarks for Language Models Abdulrahman Diao, Toluwani Aremu, Nils Lukas . The 1st Workshop on GenAI Watermarking, 2025. |
| Working Papers | | Differentially Private Inference for Large Language Models , Submitted. Rushil Thareja, Preslav Nakov, Praneeth Vepakomma, Nils Lukas . 2025. |
| | | Mitigating Watermark Forgery in Generative Models via Randomized Key Selection , Submitted. Toluwani Aremu, Noor Hazim Hussein, Munachiso S Nwadike, Samuele Poppi, Jie Zhang, Karthik Nandakumar, Neil Zhenqiang Gong, Nils Lukas . 2025. |
| | | Collaborative Threshold Watermarking , Submitted. Tameem Bakr, Anish Ambreth, Nils Lukas . 2025. |
| | | Robust and Calibrated Detection of Authentic Multimedia Content , Submitted. Sarim Hashmi, Abdelrahman Elsayed, Mohammed Talha Alam, Samuele Poppi, Nils Lukas . 2025. |
| Research Talks | Adaptively Robust and Forgery-Resistant Watermarking | 2025 |
| | ▪ Meta (FAIR), hosted by Hady Elsahar | |
| | Optimizing Adaptive Attacks against Content Watermarks | 2024 |
| | ▪ DeepMind, hosted by David Stutz | |
| | ▪ University of California, Berkeley, hosted by Dawn Song | 2024 |
| | Analyzing Leakage of Personal Information in Language Models | |
| | ▪ Microsoft M365, hosted by Robert Sim | 2024 |
| Keynotes | ▪ Meta, hosted by Will Bullock | 2023 |
| | ▪ MongoDB, hosted by Marilyn George and Archita Agarwal | 2023 |
| | How Reliable is Watermarking for Image Generators? | |
| | ▪ Google, hosted by Somesh Jha | 2023 |
| | ▪ University of California, Berkely, hosted by Dawn Song | 2023 |
| | Aviation Future Week , hosted by Emirates, Dubai | 2024 |
| | Cyber Energy Leadership Forum , Abu Dhabi | 2024 |
| Work Experience | Assistant Professor , MBZUAI, Abu Dhabi, UAE | since 08/2024 |
| | Research Intern , Royal Bank of Canada, Borealis AI, Toronto | 2024 |
| | ▪ Vertical Federated Learning, hosted by Kevin Wilson | |
| | Research Intern , Microsoft Research, Cambridge, UK | 2022 |
| | ▪ Privacy for Language Models, hosted by Shruti Tople & Lukas Wutschitz | |
| | Research Assistant , RWTH-Aachen, Aachen | 2014 - 2018 |
| | Student Researcher , DSA Daten- und Systemtechnik GmbH, Aachen | 2016 |
| | Software Engineer Intern , A.R. Bayer DSP Systeme GmbH, Düsseldorf | 2012 |

| | | |
|-----------------|---|------------------|
| Research Grants | Awarded | |
| | ▪ [Etihad Airways] Conversational Booking Agents. PI: Nils Lukas , Co-PIs: Salem Lahlou, Alham Fikri, Martin Takac, Mingming Gong [450 000 USD] | 2025 |
| | ▪ [United Al-Saqer Group] Privacy-preserving Brain Computer Interfaces. PI: Abdulrahman Mahmoud, Co-PIs: Nils Lukas , Elizabeth Churchill [136 000 USD] | 2025 |
| | Selected | |
| | ▪ [TII Funding] GFlowNets for Fuzzing of Agentic Applications. Salem Lahlou & Nils Lukas [136 000 USD] | 2025 |
| Teaching | ▪ [Amazon Special Call] AdvSim2Real: Simulating Adversarial Environments (EOI shortlisted), PI: Nils Lukas , Co-PI: Praneeth Vepakomma [100 000 USD] | 2025 |
| | Instructor , MBZUAI, UAE | |
| | ▪ ML8502: Machine Learning Security (14 weeks) | 2025 |
| | ▪ ML807: Federated Learning (7 weeks) | 2025 |
| | ▪ ML818: Emerging Topics in Trustworthy Machine Learning (4 weeks) | 2024 |
| Service | Teaching Assistant , University of Waterloo, Canada | |
| | ▪ CS458/658: Computer Security and Privacy | 2020, 2021 |
| | ▪ CS246 - Object Oriented Programming | 2021 |
| | Co-Instructor , RWTH-Aachen, Germany | |
| | ▪ Course: Data-driven Medicine | 2018 |
| Service | Area Chair | |
| | ▪ International Conference on Learning Representations (ICLR) | 2026 |
| | ▪ International Conference on Machine Learning (ICML) | 2026 |
| | Program Committee | |
| | ▪ ACM Conference on Computer and Communications Security (CCS) | 2025 |
| | ▪ IEEE Symposium on Security and Privacy (IEEE S&P) | 2025, 2026 |
| | ▪ Recent Advances in Intrusion Detection (RAID) | 2024 |
| | Artifact Evaluation Committee | |
| | ▪ The ACM Conference on Computer and Communications Security (CCS) | 2023, 2024 |
| | Reviewer | |
| | ▪ NETYS | 2025 |
| | ▪ ACM TheWebConf (WWW) | 2025 |
| | ▪ International Conference on Learning Representations (ICLR) | 2024, 2025 |
| | ▪ International World Wide Web Conference (TheWebConf) | 2024 |
| | ▪ Recent Advances in Intrusion Detection (RAID) | 2023 |
| | ▪ Neural Information Processing Systems (NeurIPS) | 2022, 2023 |
| | ▪ International Conference on Machine Learning (ICML) | 2022, 2025 |
| | ▪ The Conference on Information and Knowledge Management (CIKM) | 2020 |
| | Other | |
| | ▪ Sub-Reviewer , Proceedings on Privacy Enhancing Technologies (PETS) | 2021, 2022, 2023 |
| | ▪ Session Chair , IEEE Symposium on Security and Privacy (S&P) | 2023 |
| | ▪ Organization , Workshop on Semantic Web Solutions for Large-Scale Biomedical Data Analytics (SeWeBMeDA) | 2018 |
| | ▪ Chair for the invited faculty talk program , International Symposium on Trustworthy Foundation Models at MBZUAI | 2025 |
| | ▪ Faculty Search Committee , Machine Learning Department at MBZUAI | 2025 |
| | ▪ Admission's Committee , MBZUAI Machine Learning Department | 2025 |
| | Student Board Member , Cybersecurity and Privacy Institute | 2022, 2023, 2024 |
| | School Advisory Committee on Appointments Liaison , CrySP Lab | 2022 |