Martin Gubri

Researcher in Trustworthy AI

Tübingen, Germany
martin[at]gubri[dot]eu
gubri.eu

Research Interests

My research focuses on **auditing black-box AI systems** to make AI safer, more transparent, and accountable. Through external testing of large language models and computer vision models, I develop methods to independently assess risks and ensure compliance without access to model internals. I am now extending this programme to a **decentralised ecosystem of models**, where trust, supply chain integrity, and system-level safety are harder to guarantee.

Keywords Trustworthy AI, AI Safety, Language Models, Privacy, Security, Adversarial Machine Learning, Machine Learning Auditing, Uncertainty, Evaluation, LLM Agent, AI Governance.

Academic Positions

Since 2023 **Research Lead**, *Parameter Lab*, Tübingen, Germany.
Supervision of a team of 1–6 research interns on trustworthy AI for LLMs. Advisor: Seong Joon Oh.

2019–2023 **Doctoral Researcher**, *University of Luxembourg*, Luxembourg.
PhD on adversarial machine learning.

Education

2019–2023 **PhD in Computer Science**, *University of Luxembourg*, Luxembourg.
Thesis: *What Matters in Model Training to Transfer Adversarial Examples*.
Supervisors: Yves Le Traon, Maxime Cordy. In collaboration with *University of California, Berkeley*.
Funded by FNR CORE, STELLAR Testing Self-Learning Systems.

2014–2015 **Specialised Master's in Data Science**, *ENSAE Paris*, France.
Graduated with high honour.

2012–2014 **Dual Master's Degrees**, *Toulouse School of Economics*, France.
Master's in Statistics and Econometrics (rank 1/28)
Magister in Economics and Statistics (rank 3/16, joint with Paul Sabatier University).

2011–2012 **Bachelor's in Economics and Mathematics**, *Toulouse School of Economics*, France.

2009–2011 **Classe préparatoire B/L**, *Lycée Montaigne*, France.
Humanities, social sciences and mathematics.

Awards & Distinctions

2026 **Best paper award**, ICLR-CAO'26 workshop (DISCO). **Acceptance rate: 0.8%** (1/126)

2023 **Spotlight** paper, NeurIPS'23 (ProPILE). **Acceptance rate: 3.1%**

Publications

Overview

- Citations: 730 · h-index: 10 · i10-index: 10 (Google Scholar, May 2026)
- Publications (CORE): 7 A* main · 3 A* other tracks/findings · 1 A main · 1 A findings
- Journals (JCR): 1 Q1

In Conference Proceedings

2026 A. Rubinstein, B. Raible, **M. Gubri**, and S. J. Oh. DISCO: Diversified sample condensation for accelerating model evaluation. In ***ICLR*** *& ICLR-CAO (****Best Paper Award****)*, 2026. [link].

2026 A. Heakl, **M. Gubri**, S. Khan, S. Yun, and S. J. Oh. Dr.LLM: Dynamic layer routing in LLMs. In ***ICLR***, 2026. [link].

2026 A. Goel, C. Emde, S. J. Oh, S. Yun, and **M. Gubri**. Privacy collapse: Benign fine-tuning can break contextual privacy in language models. In ***ACL***, 2026. Acceptance rate: 19%. [link].

2026 C. Emde, A. Rubinstein, A. Goel, A. Heakl, S. Yun, S. J. Oh, and **M. Gubri**. MASEval: Extending multi-agent evaluation from models to systems. In ***ACL Demo***, 2026. [link].

2025 H. Puerto, **M. Gubri**, S. Yun, and S. J. Oh. Scaling up membership inference: When and how attacks succeed on large language models. In ***NAACL Findings***, 2025. [link].

2025 H. Puerto, **M. Gubri**, T. Green, S. J. Oh, and S. Yun. C-SEO Bench: Does conversational SEO work? In ***NeurIPS D&B*** *Track*, 2025. [link].

2025 T. Green, **M. Gubri**, H. Puerto, S. Yun, and S. J. Oh. Leaky thoughts: Large reasoning models are not private thinkers. In ***EMNLP***, 2025. [link].

2024 D. Ulmer, **M. Gubri**, H. Lee, S. Yun, and S. Oh. Calibrating large language models using their generations only. In ***ACL***, 2024. [link].

2024 **M. Gubri**, D. Ulmer, H. Lee, S. Yun, and S. J. Oh. TRAP: Targeted random adversarial prompt honeypot for black-box identification. In ***ACL Findings***, 2024. [link].

2023 S. Kim, S. Yun, H. Lee, **M. Gubri**, S. Yoon, and S. J. Oh. ProPILE: Probing privacy leakage in large language models. In ***NeurIPS (spotlight)***, 2023. [link].

2022 **M. Gubri**, M. Cordy, M. Papadakis, Y. Le Traon, and K. Sen. LGV: Boosting adversarial example transferability from large geometric vicinity. In ***ECCV***, 2022. [link].

2022 **M. Gubri**, M. Cordy, M. Papadakis, Y. Le Traon, and K. Sen. Efficient and transferable adversarial examples from Bayesian neural networks. In ***UAI***, 2022. [link].

2022 A. Franci, M. Cordy, **M. Gubri**, M. Papadakis, and Y. L. Traon. Influence-driven data poisoning in graph-based semi-supervised classifiers. In *CAIN*, 2022. [link].

2020 S. Ghamizi, M. Cordy, **M. Gubri**, M. Papadakis, A. Boystov, Y. Le Traon, and A. Goujon. Search-based adversarial testing and improvement of constrained credit scoring systems. In ***ESEC/FSE***, 2020. [link].

Journal Articles

2026 O. Zeyen, M. Cordy, **M. Gubri**, G. Perrouin, and M. Acher. Testing uniform random samplers: Methods, datasets and protocols. ***ACM TOSEM***, 2026. [link].

Workshop Papers

2026 A. Mohamed and **M. Gubri**. Is multilingual LLM watermarking truly multilingual? Scaling robustness to 100+ languages via back-translation. In ***ICML****-TAIGR*, 2026. [link].

2026 C. Emde, A. Goel, S. Yun, S. J. Oh, and **M. Gubri**. Lost in communication: Uncertainty propagation in multi-agent systems. In ***ICML****-AgenticUQ*, 2026. [link].

2025 A. Davies, E. Nguyen, M. Simeone, E. Johnston, and **M. Gubri**. Position: Social science is necessary for operationalizing socially responsible foundation models. In ***ICLR****-HAIC*, 2025. [link].

Unpublished (Preprints and Under Submission)

2026 E. S. Ruzzetti, C. Emde, S. Yun, S. J. Oh, and **M. Gubri**. MuPPET: A benchmark for contextual privacy of LLM assistants in multi-party conversations, 2026. [link].

2023 **M. Gubri**, M. Cordy, and Y. L. Traon. Going further: Flatness at the rescue of early stopping for adversarial example transferability, 2023. [link].

2018 **M. Gubri**. Adversarial perturbation intensity achieving chosen intra-technique transferability level for logistic regression, 2018. [link].

Research Funding

Industry Sponsored

2025 Coauthor of a research contract with Naver, funding one research scientist and five research interns (amount confidential).

Cloud Credits

2024 Google for Startups Cloud Program, 25k USD.

2024–2025 AWS Activate, 25k USD.

Research Supervision

PhD Students

2025–2026 Elena Sofia Ruzzetti, University of Rome Tor Vergata, 4-month internship at Parameter Lab.

2025 Anmol Goel, TU Darmstadt, 3-month internship at Parameter Lab.
Paper: *Privacy Collapse* (ACL 2026).

2025 Cornelius Emde, University of Oxford, 4-month internship at Parameter Lab.
Software: MASEval (ACL Demo 2026) ; Paper: ongoing.

2025 Tommaso Green, University of Mannheim, 4-month internship at Parameter Lab.
Paper: *Leaky Thoughts* (EMNLP 2025).

2024–2025 Haritz Puerto, TU Darmstadt, 8-month internship at Parameter Lab.
Papers: *Scaling Up Membership Inference* (NAACL Findings 2025), *C-SEO Bench* (NeurIPS D&B 2025).

2023 Dennis Ulmer, IT University of Copenhagen, 4-month internship at Parameter Lab.
Paper: *Apricot* (ACL 2024).

Master Students

2025 Ahmed Heakl, MBZUAI, 4-month internship at Parameter Lab.
Paper: *Dr.LLM* (ICLR 2026).

2025 Asim Mohamed, African Institute for Mathematical Sciences, Master thesis.
Fully independent supervision. Paper: *STEAM*, under review.

2021 Adriano Franci, University of Luxembourg, Master thesis.
Paper: *Semi-Supervised Data Poisoning* (CAIN 2022).

Teaching

2023–2025 Master: Cybersecurity & AI, University of Luxembourg
Authored, taught and graded one session on adversarial attacks against LLMs (course and practical). Lecturer: Maxime Cordy

2025 Bachelor: Low-techisation and Digital Technology (IS03), University of Technology of Compiègne
Mentored project work for the course tutorials. Lecturer: Stéphane Crozat

2022 Master: Advanced Topics in Applied Machine Learning, University of Luxembourg
Authored and taught two sessions, co-organized the course, and co-authored the final project. Lecturer: Mike Papadakis

2022 Master: Introduction to Machine Learning, University of Luxembourg
Authored and taught six sessions. Authored the exam. Lecturer: Mike Papadakis

2022 PhD: ML Security in the Real World, Cyberwal Doctoral Winter School, Belgium
Co-authored and taught the practical component of the tutorial. Lecturer: Maxime Cordy

2021 Master: Introduction to Machine Learning, University of Luxembourg
Authored and taught two introductory sessions to ML. Lecturer: Yves Le Traon

2020 Bachelor: Software Engineering 2, University of Luxembourg
Authored and taught four sessions on ML engineering. Lecturer: Yves Le Traon

Invited Talks

2026 Auditing Black-Box LLMs: An Investigator's Toolkit for Security and Safety
École Polytechnique (France), Télécom Paris (France), Idiap (Switzerland) [slides]

2025 Revealing the Invisible: Auditing the Hidden Risks of Black-Box LLMs
University of Mannheim (Germany), University of Trento (Italy), University of Luxembourg (Luxembourg)

2024 Trustworthy Machine Learning in the Era of Large Language Models, CENIA (Chile) [slides]

Academic Services

Senior Role

Area Chair UAI 2026.

Reviewer

NeurIPS 2026, ACL ARR (May 2026 cycle for EMNLP 2026, Mar. 2026 cycle, Jan. 2026 cycle for ACL 2026, May 2025 cycle for EMNLP 2025), UAI 2024, UAI 2023, International Journal of Computer Vision, IEEE Transactions on Neural Networks and Learning Systems, IEEE Transactions on Pattern Analysis and Machine Intelligence, Software Testing Verification and Reliability (Wiley), SiMLA Workshop 2023.

Professional Memberships

ACL (since 2024), ACM (since 2026)

Additional Experience

Since 2012 **Board Member (previously Co-President)**, *Framasoft*, France.
Non-profit governance of a 10-employee organisation promoting digital commons.

2017–2019 **Freelance Data Scientist**, France.
Industrial projects on machine learning, sampling design, data visualization & web scraping.

2015–2016 **Ford–Mozilla Technology Exchange Fellow**, *ONG Derechos Digitales*, Chile.
Developer and technology referent on security, privacy, cryptography & net neutrality for the NGO.

2015 **Google Summer of Code**, *R Foundation*.
Development for spdep, the spatial statistics R package. Supervised by Roger Bivand & Giovanni Millo.

2014–2015 **R&D Scientist**, *BIwhere consulting*, Paris.
Development of applied spatial econometric methods.

2013 **Creator & Developer**, *Measure Net Neutrality*, La Quadrature du Net.
Open-source tool measuring net neutrality. *Winner of the Open World Forum Student DemoCup*, 2014.

Technical Skills and Contributions

Open-Source Contributions

Co-Creator MASEval: Supervised the development of a multi-agent system evaluation library.

Security 30+ vulnerability disclosures in open-source software (2017–2019, list on gubri.eu).

Significant Torchattacks, Adversarial Robustness Toolbox (ART), spdep R package, Mozilla Common Voice.

Minor Huggingface's transformers, NVIDIA's TensorRT-Model-Optimizer, scikit-learn, GCG, i.a.

Technical Skills

ML & DL PyTorch, TensorFlow (basic), scikit-learn, Statsmodels, R Stats Package.

Programming Python, R (incl. tidyverse), Bash, LaTeX.

Languages

English Fluent
French Native
Spanish Fluent
German Basic

Outreach

2025 **Media Coverage**.
- *Privacy Collapse* featured by AI World among the **top AI papers of the week**.
- *Leaky Thoughts* featured by DAIR.AI and The AI Timeline among the **top AI papers of the week**.
- *Dr.LLM* featured by DAIR.AI among the **top AI papers of the week**.

2025 **Scientific Advisor**, *AI Public Outreach*, Framasoft.
Provided scientific guidance on AI for Framasoft's governance and public-facing projects, including FramamIA and Lokas.

2023 **Public Demo**, *ProPILE Paper*, Parameter Lab.
Lets individuals assess whether their personal data was memorised by commercial LLMs. [link]

2017 **Public Talk**, *Capitole du Libre*, Toulouse, France.
XSS and Free Software: Houston, we have a problem (in French).

2016 **Public Talk**, *Primavera Hacker*, Santiago, Chile.
What does the BIP! card know about you? Risk analysis & practical demonstration (in Spanish). National press coverage in *Las Últimas Noticias* [1] [2].

Referees

Dr. Seong Joon Oh
Associate Professor
KAIST
✉ coallaoh[at]gmail.com
seongjoonoh.com

Prof. Dr. Yves Le Traon
Full Professor & Director of the Interdisciplinary Centre for Security, Reliability and Trust (SnT)
University of Luxembourg
✉ yves.letraon[at]uni.lu

Dr. Sangdoo Yun
Research Director
Naver AI Lab
✉ sangdoo.yun[at]navercorp.com
sangdooyun.github.io

Dr. Maxime Cordy
Assistant Professor
University of Luxembourg
✉ maxime.cordy[at]uni.lu
maxcordy.github.io

Dr. Hwaran Lee
Assistant Professor
Sogang University
✉ hwaranlee[at]sogang.ac.kr
hwaranlee.github.io

Dr. Mike Papadakis
Associate Professor
University of Luxembourg
✉ michail.papadakis[at]uni.lu
mpapad.github.io