Edoardo **Debenedetti**

PhD Student in Computer Science @ ETH Zürich

(+41) 76 699 43 27 | edebenedetti@inf.ethz.ch | https://edoardo.science | dedeswim | Edoardo Debenedetti

Education

ETH Zürich - Federal Institute of Technology Zürich

Zürich, Switzerland

PhD in Computer Science — *08/2022 - Q1/2026 (exp.)*

- Focus: **Real-world machine learning security and privacy**, advised by **Prof. Florian Tramèr** in the **SPY Lab**.
- **IT Coordinator** for the group: managing the GPU servers and hardware resources.

EPFL - Federal Institute of Technology Lausanne

Lausanne, Switzerland

MSc in Computer Science — *09/2019 - 04/2022*

- **GPA 5.63/6**, focus on **Machine Learning, Security, and Privacy**.
- Master's Thesis about the **adversarial robustness of Vision Transformers** supervised by **Princeton University**'s **Prof. Mittal**.

Politecnico di Torino

Turin, Italy

BSc in Computer Engineering — *09/2016 - 07/2019*

- **GPA 28.4/30**, graduation mark 110/110, **top 9%**.
- **Exchange year at 同济大学** (Tongji University), in Shanghai (China).

Industry experience

Meta

Menlo Park, CA, United States

Research Scientist Intern — *07/2025 - 11/2025*

- In the GenAI Red Team, working on agents security.

Google

Munich, Germany – Zurich, Switzerland

Student Researcher — *10/2024 - 02/2025*

- Worked on CaMeL, a system level defense against prompt injection attacks, to build secure AI agents.
- Co-hosted by Tianqi Fan (Google ML Red Team) and Ilia Shumailov (Google DeepMind).

Bloomberg LP

London, United Kingdom

Software Engineering Intern — *07/2021 - 09/2021*

- Worked in the **Multi Asset Risk System** team, on the re-design and implementation of the configuration of a distributed logging library.
- Move the configuration of a **distributed logging library** from an internal technology to a **centralized SQL DB**, using a **cache** and a **C++ service**.
- The configuration is checked **~1M times per minute**, and the usage of the cache gave a **~23x speed improvement** w.r.t. querying the DB.

Armasuisse Cyber-Defence Campus

Lausanne, Switzerland

Research Intern — *08/2020 - 01/2021*

- Worked on **Machine Unlearning** and **Membership Inference Attacks** against Generative Models, supervised by **Prof. Mathias Humbert**.
- Adapt the **MIA** technique proposed by the *GAN-Leaks* work (by Chen et al.), to work after the removal some datapoints from the training set.
- The technique achieved **promising results** when attacking DCGAN trained on the CelebA dataset

Publications

* denotes equal contribution.

Conference proceedings

- Carlini, N., Rando, J., **Debenedetti, E.**, Nasr, M., Tramèr, F., *"AutoAdvExBench: Benchmarking Autonomous Exploitation of Adversarial Example Defenses"*, Forty-Second International Conference on Machine Learning, 2025, **Oral**.
- Aerni, M., Rando, J., **Debenedetti, E.**, Carlini, N., Ippolito, D. Tramèr, F., *"Measuring Non-Adversarial Reproduction of Training Data in Large Language Models"*, Thirteenth International Conference on Learning Representations, 2025.
- Nestaas, F., **Debenedetti, E.**, Tramèr, F., *"Adversarial Search Engine Optimization for Large Language Models"*, Thirteenth International Conference on Learning Representations, 2025.
- **Debenedetti, E.**, Zhang, J., Balunović, M., Beurer-Kellner, L., Fischer, M. Tramèr, F., *"AgentDojo: A Dynamic Environment to Evaluate Attacks and Defenses for LLM Agents"*, Thirty-eighth Conference on Neural Information Processing Systems Datasets and Benchmarks Track, 2024. (**SafeBench First Prize**).
- **Debenedetti, E.***, Rando, J.*, Paleka, D.*, Silaghi, F., Albastroiu, D., Cohen, N., Lemberg, Y., Ghosh, R., Wen, R., Salem, A., Cherubin, G., Zanella-Beguelin, S., Schmid, R., Klemm, V., Miki, T., Li, C. Kraft, S., Fritz, M., Tramèr, F., Abdelnabi, S., Schönherr, L. *"Dataset and Lessons Learned from the 2024 SaTML LLM Capture-the-Flag Competition"*, Thirty-eighth Conference on Neural Information Processing Systems Datasets and Benchmarks Track, 2024 (**Spotlight**).
- Chao, P.*, **Debenedetti, E.***, Robey, A.*, Andriushchenko, M.*, Croce, F., Sehwag, V., Dobriban, E., Flammarion, N., Pappas, G., Tramèr, F, Hassani, H., Wong, E., *"JailbreakBench: An Open Robustness Benchmark for Jailbreaking Language Models"*, Thirty-eighth Conference on Neural Information Processing Systems Datasets and Benchmarks Track, 2024.
- **Debenedetti, E.**, Severi, G., Carlini, N., Choquette-Choo, C. A., Jagielski, M., Nasr, M., Wallace, E., Tramèr, F., *"Privacy Side Channels in Machine Learning Systems"*, 33rd USENIX Security Symposium, 2024.
- **Debenedetti, E.**, Carlini, N., Tramèr, F., *"Evading Black-box Classifiers Without Breaking Eggs"*, 2nd IEEE Conference on Secure and Trustworthy Machine Learning, 2024, **Distinguished Paper Award Runner-up**.
- **Debenedetti, E.**, Sehwag, V., Mittal, P., *"A Light Recipe to Train Robust Vision Transformers"*, 1st IEEE Conference on Secure and Trustworthy Machine Learning, 2023.
- Croce, F.*, Andriushchenko, M.*, Sehwag, V.*, **Debenedetti, E.***, Flammarion, N., Chiang, M., Mittal, P., Hein, M., *"RobustBench: a standardized adversarial robustness benchmark"*, Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track, 2021.

Workshop papers

- Freeman, J., Rippe, C., **Debenedetti, E.**, Andriushchenko, M., *"Exploring Memorization and Copyright Violation in Frontier LLMs: A Study of the New York Times v. OpenAI 2023 Lawsuit"*, NeurIPS Safe Generative AI Workshop, 2024.
- **Debenedetti, E.**, Wan, Z., Andriushchenko, M., Sehwag, V., Bhardwaj, K, Kailkhura, B., *"Scaling Compute Is Not All You Need for Adversarial Robustness"*, ICLR Workshop on Reliable and Responsible Foundation Models, 2024.

Manuscripts

- Beurer-Kellner, L., Buesser, B., Creţu, A., **Debenedetti, E.**, Dobos, D., Fabian, D., Fischer, M., Froelicher, D., Grosse, K., Naeff, D., Ozoani, E., Paverd, E., Tramèr, F, Volhejn, V., *"Design Patterns for Securing LLM Agents against Prompt Injections"*, arXiv ePrint 2506.08837, 2025 ($\alpha\beta$ order).
- Carlini, N., Nasr, M., **Debenedetti, E.**, Wang, B., Choquette-Choo C., Ippolito, D., Tramèr, F., Jagielski, M., *"LLMs unlock new paths to monetizing exploits"*, arXiv ePrint 2505.11449, 2025.
- **Debenedetti, E.**, Shumailov, I., Fan, T., Hayes, J., Carlini, N., Fabian, D., Kern, C., Shi, C., Terzis, A., Tramèr, F., *"Defeating Prompt Injections by Design"*, arXiv ePrint 2503.18813, 2025.
- Qi, X., Huang, Y., Zeng, Y., **Debenedetti, E.**, Geiping, J., He, L., Huang, K., Madhushani Sehwag, U., Sehwag, V., Shi, W., Wei, B., Xie, T., Chen, D., Chen, P., Ding, J., Jia, R., Ma, J., Narayanan, A., Su, W., Wang, M., Xiao, C., Li, B., Song, D., Henderson, P., Mittal, P., *"AI Risk Management Should Incorporate Both Safety and Security"*, arXiv ePrint 2405.19524, 2024.

Honors and Awards

2025 **SafeBench First Prize,** 50'000 USD prize for the AgentDojo Benchmark.

2025 **Oral Acceptance — ICML 2025,** Top 0.98% of submitted papers.

2024 **Spotlight Acceptance — NeurIPS 2024 Datasets and Benchmarks Track,** Top 4% of submitted papers.

2024 **Distinguished Paper Award Runner-up — IEEE SaTML,** Top 1% of submitted papers.

2023 **Oral presentation — ICML AdvML Frontiers Wokshop,** Top 10% accepted papers.

2023 **CYD Doctoral Fellowship,** full PhD funding for 4 years, worth **USD 536'000** (CHF 461'000), from Armasuisse CYD Campus and EPFL. Only used for ∼1.5 years.

2021 **Google TPU Research Cloud Program,** extensive **hardware support for 8 months** to work on the Master's Thesis.

2021 **Best Paper Honorable Mention — ICLR Workshop on Security and Safety in ML Systems,** top 2 out of 50 accepted papers.

Invited talks

- **armasuisse Cyber-Alp Retreat.** – *Defeating Prompt Injections by Design.*, 2025.
- **ICLR Workshop on Building Trust in LLMs and LLM Applications.** – *Evaluating and Defending against Prompt Injection Attacks*, 2025.
- **UMass AI Security and Privacy Seminar** – *Defeating Prompt Injections by Design*, 2025.
- **Google AE Summit** – *Prompt Injection Attacks: A Critical Risk for Deployed AI Agents*, 2025.
- **Google** – *Defeating Prompt Injections by Design*, 2025.
- **Google** – *AgentDojo: A Dynamic Environment to Evaluate Attacks and Defenses for LLM Agents*, 2024.
- **Princeton Language and Intelligence** – *AgentDojo: A Dynamic Environment to Evaluate Attacks and Defenses for LLM Agents*, 2024.
- **armasuisse Cyber-Alp Retreat** – *Evading Black-box Classifiers Without Breaking Eggs*, 2024.
- **ACL SIGSEC** – *Privacy Side-channels in Machine Learning Systems*, 2023.
- **TU Graz EfficientML Reading Group** – *Privacy Side-channels in Machine Learning Systems*, 2023.

Press coverage

- **Ars Technica** – *"Researchers claim breakthrough in fight against AI's frustrating security hole"*, April 2025.
- **MIT Technology Review** – *"Cyberattacks by AI agents are coming"*, April 2025.

Teaching

- **Privacing Enhancing Technologies** – ETH Zürich: 2024 (Teaching Assistant)
- **Information Security Lab** – ETH Zürich: 2022, 2023 (Teaching Assistant)
- **Large Language Models** – ETH Zürich: 2023, 2024, 2025 (Teaching Assistant)

Professional Service

Reviewer

- **ICLR**: 2025
- **NeurIPS**: 2024, 2025
- **NeurIPS Datasets and Benchmarks Track**: 2022, 2023, 2024
- **CCS AISec workshop**: 2023, 2024

Conference service

- **Competition organizer at SaTML 2024**: lead organizer of the *Large Language Models Capture-the-Flag*. More than **400 users and 140 teams** signed up and more than 70 defenses were submitted. The competition report was accepted at NeurIPS 2024 and awarded a spotlight.
- **Volunteer at NeurIPS 2021**: helped with monitoring the website and technical issues.

Open Source Maintainer

- **AgentDojo**: Benchmark for Prompt Injection Attacks and Defenses.
  - Lead the development of the environment and wrote the documentation.
  - **239 stars** (measured in August 2025).

  Repository at `https://github.com/ethz-spylab/agentdojo`.
- **RobustBench**: adversarial robustness benchmarking library and model zoo.
  - More than 150 models spanning 3 datasets and 3 threat models.
  - **730 stars** (measured in August 2025).
  - Refactored the code to improve the extensibility of the library.

  Repository at `https://github.com/RobustBench/robustbench`.

Last update August 20, 2025.