

# Maksym Andriushchenko – Curriculum Vitae

## Personal Data

**Site:** <https://andriushchenko.me/>  
**Email:** [maksym@andriushchenko.me](mailto:maksym@andriushchenko.me)

**Scholar:** <https://scholar.google.com/citations?user=ZNtuJYoAAAAJ>  
**Github:** <https://github.com/max-andr/>

## Education

**École Polytechnique Fédérale de Lausanne (EPFL), Switzerland** (*Sep 2019 - Oct 2024*)  
PhD Degree in Computer Science advised by Nicolas Flammarion — *awarded with the Google PhD Fellowship, Open Phil AI PhD fellowship, Patrick Denantes Memorial Prize for the best thesis in the CS department of EPFL*

**Saarland University, Germany** (*Oct 2016 – Aug 2019*)  
Master's Degree in Computer Science advised by Matthias Hein from the University of Tübingen

**Dnipro National University of Railway Transport, Ukraine** (*Sep 2012 – June 2016*)  
Bachelor's Degree in Software Engineering — *with honors*

## Work

**EPFL**      **Time:** September 2019 – now  
**Role:** PhD student → postdoctoral researcher supervised by Nicolas Flammarion.

**Gray Swan AI**      **Time:** February 2024 – November 2024 (part-time consultant)  
**Role:** Member of technical staff working with leading AI safety organizations (Anthropic, UK AI Safety Institute, Center for AI Safety)

**Adobe Research**      **Time:** July 2021 – October 2021  
**Role:** Research internship supervised by John Collomosse

**PrivatBank**      **Time:** November 2015 – June 2016 (part-time)  
**Role:** Data scientist working on predictive modeling and e-commerce applications

## Awards and Grants

**Main awards**      **Patrick Denantes Memorial Prize for the best PhD thesis in the School of Computer and Communication Sciences at EPFL (\$5k award)**  
G-Research PhD Thesis Prize in Quantitative Research (2nd place, \$3k award)  
**Google PhD fellowship 2022-2025 (\$80k per year for 3 years)**  
**Open Philanthropy AI PhD Fellowship 2022-2024 (\$10k per year for travel/equipment)**  
EDIC PhD fellowship from EPFL for the first year (\$60k)  
DAAD MSc scholarship for 2 years to study at Saarland University (\$20k)

**Co-authored grants**      Safe GenAI via Robust Content Moderation Models (\$100k funded by Google in 2024)  
Safety, Robustness, and Alignment of LLM agents (\$40k funded by Google in 2024)  
OpenAI Researcher Access Program (\$5k in API credits in 2024)  
Google Research Collab 2022-2023 (\$80k for one year + \$20k in cloud compute)

**Awards for papers and competitions**      [Trojan detection challenge](#) at SatML'24: **first place**  
[Swiss AI Safety Prize](#) (2024): **award for one of the top paper submissions**  
Joint Conference of Korean AI Association (2023): **best paper award**  
ICLR Workshop on Security & Safety in ML Systems (2021): **best paper honorable mention**  
Swiss Machine Learning Day (2019): **best paper award**

## Academic Service

**Participant**      Red teaming of the **OpenAI fine-tuning** service as an external expert (October 2023)  
[Robust AI 4-day workshop](#) organized by Airbus AI Research and TNO (January 2021)

**Reviewer**      TMLR, ICLR'25, NeurIPS'24, NeurIPS'23, ICML'23, NeurIPS'22 (**top reviewer**), ICML'22, NeurIPS'21, ICML'21, CVPR'21, ICLR'21 (**outstanding reviewer**), NeurIPS'20 (**top 10% reviewers**)

<b>Workshop organization</b>	Workshop on Secure and Robust AI Agents: Bridging Machine Learning and Cybersecurity (submitted to <b>ICLR'25</b> ; organizers: Kamilė Lukošūtė, <b>Maksym Andriushchenko</b> , Baturay Saglam, Amanda Minnich, Reza Shokri, Adam Swanda, Amin Karbasi, Yaron Singer, Shafi Goldwasser)
<b>Program committee in workshops</b>	<b>NeurIPS'24</b> Red Teaming GenAI Workshop, <b>NeurIPS'24</b> SATA Workshop: Towards Safe & Trustworthy Agents, <b>NeurIPS'24</b> 3rd Workshop on New Frontiers in Adversarial ML, <b>ICML'24</b> Workshop on the Next Generation of AI Safety, <b>NeurIPS'23</b> R0-FoMo Workshop on Robustness of Few-shot and Zero-shot Learning in Foundation Models, <b>NeurIPS'23</b> Workshop on Distribution Shifts: New Frontiers with Foundation Models, <b>ICML'23</b> 2nd ICML Workshop on New Frontiers in Adversarial ML, <b>ICLR'23</b> Workshop on Pitfalls of Limited Data and Computation for Trustworthy ML, <b>NeurIPS'22</b> Workshop on Distribution Shifts, <b>NeurIPS'22</b> ML Safety Workshop, <b>ICML'22</b> New Frontiers in Adversarial ML, <b>ICML'22</b> Principles of Distribution Shift, <b>NeurIPS'21</b> Distribution Shifts: Connecting Methods and Applications, <b>ICML'21</b> Uncertainty and Robustness in Deep Learning, <b>CVPR'21</b> Adversarial ML in Real-World Computer Vision Systems, <b>ICLR'21</b> Robust and Reliable ML in the Real World, Security and Safety in ML Systems, <b>ICML'20</b> Uncertainty and Robustness in Deep Learning, <b>CVPR'20</b> Adversarial ML in Computer Vision, <b>ICLR'20</b> Towards Trustworthy ML ( <b>best reviewer award</b> )
<b>Outreach activities</b>	National coordinator for Switzerland at <a href="#">#ScienceForUkraine</a> Coordinator for Switzerland and admission officer at the <a href="#">Ukrainian Global University</a> AI and STEM workshop at a <a href="#">summer camp</a> for displaced Ukrainian children in Romania

## Student Supervision

<b>Alexander Panfilov (University of Tübingen)</b>	<b>PhD thesis (2024-2028)</b> : A Jailbreaking Perspective on LLM Safety (co-supervised with Jonas Geiping)
<b>Joshua Freeman (ETH Zürich)</b>	<b>MSc project (2024)</b> : Exploring Memorization and Copyright Violation in Frontier Large Language Models ( <i>accepted at NeurIPS 2024 Safe Generative AI Workshop</i> )
<b>Hao Zhao (EPFL)</b>	<b>MSc thesis (2023)</b> : Long Is More for Alignment: A Simple but Tough-to-Beat Baseline for Instruction Fine-Tuning ( <i>published at ICML'24, nominated for EPFL Outstanding Master's Thesis, featured in <a href="#">MIT Technology Review China</a></i> )
<b>Hichem Hadhri (EPFL)</b>	<b>MSc project (2023)</b> : Understanding overfitting in large language models
<b>Tiberiu Musat (EPFL)</b>	<b>BSc project (2023)</b> : Investigating key components for fast optimization of deep networks
<b>Francesco d'Angelo (EPFL)</b>	<b>PhD semester project (2023)</b> : Understanding the role of weight decay in deep learning ( <i>published at NeurIPS'24</i> )
<b>Théau Vannier (EPFL)</b>	<b>MSc project (2023)</b> : Understanding the training instability of transformers
<b>Joshua Freeman (EPFL)</b>	<b>BSc project (2022, unofficial)</b> : Automatic recognition of unexploded ordnance using transfer learning
<b>Jana Vuckovic (EPFL)</b>	<b>MSc project (2022)</b> : Rethinking the relationship between sharpness and generalization ( <i>follow-up work is published at ICML'23</i> )
<b>Mehrdad Saberi (EPFL)</b>	<b>Summer internship (2021)</b> : Wasserstein adversarial training and perceptual robustness
<b>Edoardo Debenedetti (EPFL)</b>	<b>MSc project (2021)</b> : RobustBench: a standardized adversarial robustness benchmark ( <i>published at NeurIPS'21 Datasets and Benchmarks Track; Best Paper Honorable Mention Prize at <a href="#">ICLR 2021 Workshop on Security and Safety in ML Systems</a></i> )
<b>Klim Kireev (EPFL)</b>	<b>PhD semester project (2020)</b> : On the effectiveness of adversarial training against common corruptions ( <i>published at UAI'22</i> )
<b>Etienne Bonvin (EPFL)</b>	<b>MSc project (2020)</b> : Adversarial robustness of kernel methods
<b>Oriol Barbany (EPFL)</b>	<b>MSc project (2019)</b> : Affine-invariant robust training (co-advised with Sebastian Stich)

## Teaching Experience

<b>EPFL</b>	<b>Probability &amp; Statistics 2021, 2022</b> (by E. Abbé), <b>Machine Learning 2020, 2021, 2022, 2023</b> (by M. Jaggi, N. Flammarion), <b>Advanced Algorithms 2020</b> (by M. Kapralov)
-------------	--

## Selected Publications

**M. Andriushchenko**, A. Souly, M. Dziemian, D. Duenas, M. Lin, J. Wang, D. Hendrycks, A. Zou, Z. Kolter, M. Fredrikson, E. Winsor, J. Wynne, Y. Gal, X. Davies. AgentHarm: A Benchmark for Measuring Harmfulness of LLM Agents (arXiv, October 2024) [[paper](#)]

**M. Andriushchenko**, F. Croce, N. Flammarion. Jailbreaking Leading Safety-Aligned LLMs with Simple Adaptive Attacks (a short version appeared at the ICML 2024 Workshop on the Next Generation of AI Safety) [[paper](#)]

**M. Andriushchenko**, N. Flammarion. Towards Understanding Sharpness-Aware Minimization (ICML 2022) [[paper](#)]

F. Croce\*, **M. Andriushchenko\***, V. Sehwag\*, E. Debenedetti\*, N. Flammarion, M. Chiang, P. Mittal, M. Hein. RobustBench: a standardized adversarial robustness benchmark (NeurIPS 2021 Datasets and Benchmarks Track, **Best Paper Honorable Mention Prize** at [ICLR 2021 Workshop on Security and Safety in Machine Learning Systems](#)) [[paper](#)]

**M. Andriushchenko\***, F. Croce\*, N. Flammarion, M. Hein. Square Attack: a Query-Efficient Black-Box Adversarial Attack via Random Search (ECCV 2020) [[paper](#)]

## Full Publication List

J. Freeman, C. Rippe, E. Debenedetti, **M. Andriushchenko**. Exploring Memorization and Copyright Violation in Frontier LLMs: A Study of the New York Times v. OpenAI 2023 lawsuit (a short version appeared at the NeurIPS 2024 Safe Generative AI Workshop) [soon on arXiv]

**M. Andriushchenko**, A. Souly, M. Dziemian, D. Duenas, M. Lin, J. Wang, D. Hendrycks, A. Zou, Z. Kolter, M. Fredrikson, E. Winsor, J. Wynne, Y. Gal, X. Davies. AgentHarm: A Benchmark for Measuring Harmfulness of LLM Agents (arXiv, October 2024) [[paper](#)]

A. Zou, L. Phan, J. Wang, D. Duenas, M. Lin, **M. Andriushchenko**, R. Wang, Z. Kolter, M. Fredrikson, D. Hendrycks. Improving Alignment and Robustness with Short Circuiting (NeurIPS 2024) [[paper](#)]

**M. Andriushchenko\***, F. D'Angelo\*, A. Varre, N. Flammarion. Why Do We Need Weight Decay in Modern Deep Learning? (NeurIPS 2024) [[paper](#)]

P. Chao\*, E. Debenedetti\*, A. Robey\*, **M. Andriushchenko\***, F. Croce, V. Sehwag, E. Dobriban, N. Flammarion, G.J. Pappas, F. Tramèr, H. Hassani, E. Wong. JailbreakBench: An Open Robustness Benchmark for Jailbreaking Large Language Models (NeurIPS 2024 Datasets and Benchmarks Track) [[paper](#)]

**M. Andriushchenko**, N. Flammarion. Does Refusal Training in LLMs Generalize to the Past Tense? (NeurIPS 2024 Safe Generative AI Workshop, **oral presentation**) [[paper](#)]

H. Zhao, **M. Andriushchenko**, F. Croce, N. Flammarion. Is In-Context Learning Sufficient for Instruction Following in LLMs? (NeurIPS 2024 Workshop on Adaptive Foundation Models) [[paper](#)]

J. Rando, F. Croce, K. Mitka, S. Shabalin, **M. Andriushchenko**, N. Flammarion, F. Tramèr. Competition Report: Finding Universal Jailbreak Backdoors in Aligned LLMs (arXiv, April 2024) [[paper](#)]

**M. Andriushchenko**, F. Croce, N. Flammarion. Jailbreaking Leading Safety-Aligned LLMs with Simple Adaptive Attacks (a short version appeared at the ICML 2024 Workshop on the Next Generation of AI Safety) [[paper](#)]

H. Zhao, **M. Andriushchenko**, F. Croce, N. Flammarion. Long Is More for Alignment: A Simple but Tough-to-Beat Baseline for Instruction Fine-Tuning (ICML 2024) [[paper](#)]

L. Adilova, **M. Andriushchenko**, M. Kamp, A. Fischer, M. Jaggi. Layer-Wise Linear Mode Connectivity (ICLR 2024) [[paper](#)]

**M. Andriushchenko**. Adversarial Attacks on GPT-4 via Simple Random Search (December 2023) [[paper](#)]

E. Debenedetti, Z. Wan, **M. Andriushchenko**, V. Sehwag, K. Bhardwaj, B. Kailkhura. Scaling Compute Is Not All You Need for Adversarial Robustness (ICLR 2024 Workshop on Reliable and Responsible Foundation Models) [[paper](#)]

S. Shin, D. Lee, **M. Andriushchenko**, N. Lee. The Effects of Overparameterization on Sharpness-Aware Minimization: An Empirical and Theoretical Analysis (September 2023, **best paper award** at the Joint Conference of Korean Artificial Intelligence Association (2023) [[paper](#)])

**M. Andriushchenko**, D. Bahri, H. Mobahi, N. Flammarion. Sharpness-Aware Minimization Leads to Low-Rank Features (NeurIPS 2023) [[paper](#)]

K. Kireev, **M. Andriushchenko**, C. Troncoso, N. Flammarion. Transferable Adversarial Robustness for Categorical Data via Universal Robust Embeddings (NeurIPS 2023) [[paper](#)]

**M. Andriushchenko**, F. Croce, M. Müller, M. Hein, N. Flammarion. A modern look at the relationship between sharpness and generalization. (ICML 2023) [[paper](#)]

**M. Andriushchenko**, A. Varre, L. Pillaud-Vivien, N. Flammarion. SGD with large step sizes learns sparse features (ICML 2023) [[paper](#)]

K. Kireev\*, **M. Andriushchenko\***, N. Flammarion. On the effectiveness of adversarial training against common corruptions (UAI 2022, [ICLR'21 Workshop on Robust and Reliable Machine Learning in the Real World](#)) [[paper](#)]

Michael Rose, Sanita Reinsone, **Maksym Andriushchenko**, Marcin Bartosiak, Anna Bobak et al. #ScienceForUkraine: an Initiative to Support the Ukrainian Academic Community. “3 Months Since Russia’s Invasion in Ukraine”, February 26 – May 31, 2022 (SSRN, 2022) [[paper](#)]

**M. Andriushchenko**, N. Flammarion. Towards Understanding Sharpness-Aware Minimization (ICML 2022) [[paper](#)]

**M. Andriushchenko**, X. Rebecca Li, Geoffrey Oxholm, Thomas Gittings, Tu Bui, Nicolas Flammarion, John Collomosse. ARIA: Adversarially Robust Image Attribution for Content Provenance ([CVPR 2022 Workshop on Media Forensics](#)) [[paper](#)]

F. Croce, **M. Andriushchenko**, N. Singh, N. Flammarion, M. Hein. Sparse-RS: a versatile framework for query-efficient sparse black-box adversarial attacks (AAAI 2022) [[paper](#)]

F. Croce\*, **M. Andriushchenko\***, V. Sehwal\*, E. DeBenedetti\*, N. Flammarion, M. Chiang, P. Mittal, M. Hein. RobustBench: a standardized adversarial robustness benchmark (NeurIPS 2021 Datasets and Benchmarks Track, **Best Paper Honorable Mention Prize** at [ICLR 2021 Workshop on Security and Safety in Machine Learning Systems](#)) [[paper](#)]

M. Mosbach, **M. Andriushchenko**, D. Klakow. On the Stability of Fine-tuning BERT: Misconceptions, Explanations, and Strong Baselines (ICLR 2021) [[paper](#)]

**M. Andriushchenko\***, F. Croce\*, N. Flammarion, M. Hein. Square Attack: a query-efficient black-box adversarial attack via random search (ECCV 2020) [[paper](#)]

**M. Andriushchenko**, N. Flammarion. Understanding and Improving Fast Adversarial Training (NeurIPS 2020) [[paper](#)]

**M. Andriushchenko**, M. Hein. Provably Robust Boosted Decision Stumps and Trees against Adversarial Attacks (NeurIPS 2019, contributed talk at [Workshop on Machine Learning with Guarantees](#); **best paper award** at Swiss Machine Learning Day (2019)) [[paper](#)]

M. Hein, **M. Andriushchenko**, J. Bitterwolf. Why ReLU networks yield high-confidence predictions far away from the training data and how to mitigate the problem (**oral** at CVPR 2019, 5.6% acceptance rate, contributed talk at [ICML 2019 Uncertainty and Robustness in Deep Learning Workshop](#)) [[paper](#)]

F. Croce\*, **M. Andriushchenko\***, M. Hein. Provable Robustness of ReLU Networks via Maximization of Linear Regions (AISTATS 2019) [[paper](#)]

M. Mosbach\*, **M. Andriushchenko\***, T. Trost, M. Hein, D. Klakow. Logit Pairing Methods Can Fool Gradient-Based Attacks ([NeurIPS 2018 Workshop on Security in ML](#)) [[paper](#)]

M. Hein and **M. Andriushchenko**. Formal Guarantees on the Robustness of a Classifier Against Adversarial Manipulation (NeurIPS 2017) [[paper](#)]