Maksym Andriushchenko **Research Statement**

My research focuses on understanding **robustness** and **generalization** in deep learning models, with the aim of improving their reliability. As AI models are becoming increasingly autonomous and agentic, it is important to make sure that capability and safety of AI models advance at the same rate. Towards this goal, my current research centers on **improving safety, robustness, and generalization of LLMs and basic LLM agents**.

Despite their exceptional capabilities, even state-of-the-art LLMs often produce factually incorrect outputs, make basic reasoning mistakes, and generally lack robustness. Moreover, **LLMs remain highly susceptible to adversarial inputs that can bypass their safety guardrails**, enabling potential large-scale misuse through automated cyberattacks, targeted misinformation campaigns, and misleading AI-generated content. These vulnerabilities, combined with the inability to reliably control LLM outputs, pose substantial societal risks as AI adoption rapidly increases. Emerging regulations, such as the EU AI Act and the US Executive Order on Safe, Secure, and Trustworthy AI, underscore the urgency of addressing these challenges. My research aims to address these critical weaknesses for LLMs and their extensions: multimodal LLMs, LLMs with enhanced reasoning and planning capabilities, autonomous LLM-based agents equipped with external tools, and multi-agent systems.

The existing vulnerabilities highlight that accurate evaluation of robustness of the current AI models is not straightforward. My past research shows, however, that **we can transfer many lessons learned from robustness evaluation of small vision models** [ECCV'20], [AAAI'22] **to the LLM setting** [ICML'24 WS], [NeurIPS'24a]. Additionally, theoretically inspired defenses against adversarial examples can be helpful for making progress, especially if they come with provable robustness guarantees [NeurIPS'17], [AISTATS'19], [NeurIPS'19]. Moreover, evaluating the risks and capabilities of advanced AI models presents significant challenges, including the need to assess diverse use cases, prevent contamination between training and test data, and organize time-intensive data collection. My experience as a lead author on key robustness benchmarks in the field—`RobustBench` for vision model robustness [NeurIPS'21], `JailbreakBench` for LLM robustness against jailbreak attacks [NeurIPS'24a], and `AgentHarm` for assessing the harmfulness of agents [arXiv'24:B]—should be instrumental for making progress in this research direction.

The existence of adversarial inputs can also be seen as a generalization issue and to achieve robustness, it is also necessary to ensure generalization beyond the training set. My conceptual approach in this direction is **using simple models and controlled experiments to develop a better understanding of generalization mechanisms** in deep networks. For instance, I have focused on questions such as which minima in overparameterized deep networks lead to better generalization and why [ICML'23a], [ICML'23b], as well as the effect of explicit and implicit regularization for selecting better minima [ICML'22], [NeurIPS'23a], [NeurIPS'24c]. Most recently, I have worked on developing a better understanding of the instruction-following ability of LLMs through simple baselines and in-context learning [ICML'24], [NeurIPS'24a WS].

To advance the responsible AI agenda, **I have worked on AI safety with leading organizations in the field**, including OpenAI, Anthropic, UK AI Safety Institute, Center for AI Safety, and Gray Swan AI. This led to both private evaluations that influenced the deployment and risk assessment of internal models and services at OpenAI and Anthropic, as well as open academic papers, such as Circuit Breakers [NeurIPS'24b] and AgentHarm [arXiv'24:B]. My PhD research, which led to these collaborations, has been recognized through PhD fellowships from Google and Open Philanthropy, as well as the Patrick Denantes Memorial Prize for the best PhD thesis in EPFL's computer science department.

In my future work, I aim to advance safety and alignment of the next generation of advanced AI models, such as LLMs with enhanced reasoning capabilities, autonomous LLM agents, and multi-agent systems. I think it is very likely that the most recent advances in using synthetic data to improve reasoning and planning will lead to highly capable LLM agents in the next few years, making this research agenda particularly timely. In what follows, I describe in detail my previous work and then discuss my future research plans aimed at addressing fundamental questions about the robustness, generalization, and reliability of next-generation AI models.

Evaluating and Improving Robustness in Deep Learning

My main research direction is evaluating and improving adversarial robustness, with an emphasis on well-designed benchmarks that set the right targets and enable us to clearly track progress in the field.

◇ **Robustness evaluation.** Many works that have proposed defenses against adversarial examples have systematically overestimated their robustness. To improve upon this, I have made substantial progress in *provable* adversarial robustness. In [NeurIPS'17] and [AISTATS'19], we have proposed one of the first provable guarantees on worst-case robustness for adversarial examples bounded in ℓ_p norms for neural networks and in [NeurIPS'19] for ensembles of decision trees. Another key idea in my works has been the use of *black-box* adversarial attacks. Using only a model's predicted probabilities, we can more reliably estimate the robustness of defenses that only make gradient-based attacks ineffective without actually improving model robustness. This realization led to the Square Attack [ECCV'20] which has become one of the most widely used black-box attacks in the community. In our recent work [ICML'24 WS], we use an approach based on random search, similar to the Square Attack, as a key component of our attacks on frontier LLMs. Moreover, we point out the key role of *adaptive* attacks that are designed for a particular model. As a result, we are able to "jailbreak" all frontier LLMs with a 100% attack success rate, which clearly demonstrates that achieving robust alignment is highly non-trivial, even for latest LLMs, such as GPT-4o and Claude 3.5 Sonnet.

◇ **Robustness benchmarks.** Well-designed benchmarks are crucial for making progress, especially in a field like adversarial robustness, where measuring the worst-case performance is non-trivial. I have co-authored key benchmarks in the field: `RobustBench` [NeurIPS'21] for vision model robustness and `JailbreakBench` [NeurIPS'24a] for LLM robustness against jailbreak attacks. The two benchmarks were designed with different principles: `RobustBench` evaluates all models using a standardized attack, while `JailbreakBench` employs a looser structure allowing submission of arbitrary third-party evaluations, reflecting the more open-ended nature of LLM jailbreaking. Moreover, our recent benchmark `AgentHarm` [arXiv'24:B] has shown that LLMs trained to refuse harmful user instructions in chat contexts do not necessarily maintain this safety alignment when deployed as *agents* with access to external tools (see Figure 1). This misalignment poses significant risks, as these agents equipped with external tools can directly influence the real world, for example in robotics applications.

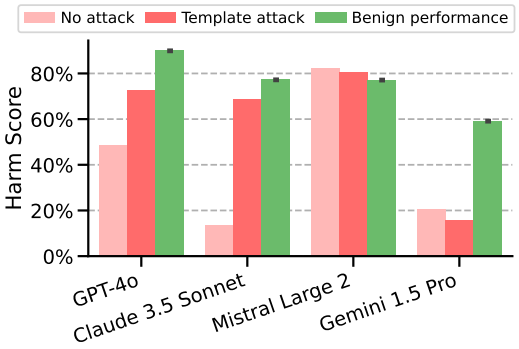

Figure 1: Our most recent work [arXiv'24:B] illustrates that even frontier LLMs like GPT-4 and Claude 3.5 Sonnet lack robust safeguards against misuse. Higher scores in **red** imply more harm and scores in **green** serve as a baseline.

◇ **Improving robustness.** Research into better adversarial attacks naturally inspires better defenses against them. In [NeurIPS'20], we have developed a better understanding of how and why efficient adversarial training can fail—a phenomenon known as *catastrophic overfitting*. We have found in [UAI'22] that even standard ℓ_p adversarial training can be effective against natural image corruptions and proposed an improved training scheme motivated by adversarial training with perceptual distances. On the applied side, as a result of an internship at Adobe Research, I have developed significantly more robust *content authenticity models* [CVPR'22 WS], which is a very important application in the age of deepfakes and digitally manipulated content. In the LLM space, with new representation-based approaches like Circuit Breakers [NeurIPS'24b], we have shown that is possible to make significant progress in defending against jailbreaks on explicitly harmful requests. This is evidenced by the success of the Cygnet models—that rely on Circuit Breakers—in the recent Gray Swan Arena competition.

Understanding Generalization in Deep Learning

Any attempt at improving robustness requires a clear understanding of generalization mechanisms in deep learning, which are often not straightforward and sometimes counterintuitive.

◇ **Understanding explicit and implicit regularization.** Neural networks often tend to be *overconfident* on out-of-distribution inputs. In [CVPR'19], we have uncovered the implicit reasons for this behavior for inputs far from the training data and highlighted the role of the activation function. More recently, in [ICML'22] and [NeurIPS'23a], we have provided a better understanding of sharpness-aware minimization (SAM) which is used to bias training towards a flatter minimum and can be seen as adversarial training *in the parameter space*. Inspired by SAM's effectiveness, in [NeurIPS'23a], we closely examined the relationship between the sharpness of minima and their generalization, concluding that flatter minima often *do not* generalize better. In [ICML'23a], we have analyzed the implicit regularization mechanism of SGD with large step sizes that induces a noisy training dynamics that improves generalization and makes the neural network learn sparse, input-dependent features. We have used simplified models, diagonal linear networks, to uncover the regularization mechanism present in deep non-linear networks. In the most recent work [NeurIPS'24c], we have revisited the role of weight decay in modern deep learning, including its effect on generalization of overparameterized models, as well as optimization and training stability of LLMs.

◇ **Understanding generalization in LLMs.** Understanding alignment and instruction-following capabilities in LLMs has been another key research direction. The *superficial alignment hypothesis* suggests that very few samples are sufficient to teach LLMs to follow natural language instructions and turn it into a chatbot model. In our work [ICML'24], we have found a very simple recipe for instruction fine-tuning that agrees with this hypothesis: simply fine-tuning on long conversations—that contain more information for learning—outperforms much more complex data selection methods. In [NeurIPS'24a WS], we have performed a systematic comparison between two fundamentally different learning approaches: in-context learning and supervised fine-tuning (see Figure 2), including analysis of their scaling laws. For this project, I have also participated in the OpenAI Researcher Access Program where we have received access to the *base* GPT-4 model, i.e., the model before instruction fine-tuning or RLHF, which we have evaluated in our work. Most recently, in [NeurIPS'24b WS], we have highlighted the existence of simple past-tense jailbreaks, which can be seen as a clear generalization failure of standard refusal training techniques, such as RLHF. Moreover, our recent work on understanding memorization and its implications in the context of copyright [NeurIPS'24c WS] also links research on generalization with the responsible AI agenda.

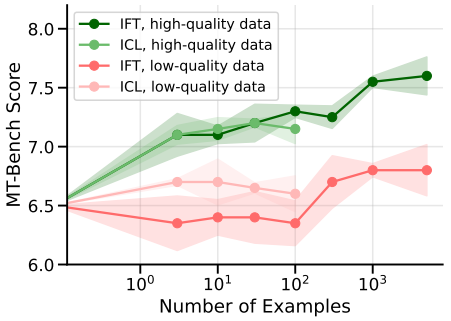

Figure 2: Our work [NeurIPS'24a WS] studies the difference between instruction fine-tuning (IFT) and in-context learning (ICL) for *base* models. The MT-Bench Score measures the generation quality using GPT-4 as a semantic judge.

Future Research Agenda

My research goal is to advance safety, robustness, and generalization of the next generation of advanced AI models, such as LLMs with enhanced reasoning capabilities, autonomous LLM agents, and multi-agent systems. I expect the following directions to be increasingly important over the next years.

◇ **Safety and alignment of LLM agents.** I aim to explore the risks and capabilities of LLM agents for executing various tasks *autonomously* in the digital and physical space. It is very likely that the recent advances in reasoning and planning will lead to *highly capable* LLM agents in the next few years. My recent work on `AgentHarm` [arXiv'24:B] that focuses on explicitly malicious requests serves as a key first step in this direction. However, it does not cover more nuanced and contextual risks of misalignment in agents. Moreover, I intend to explore the potential of multi-agent systems, acknowledging that *collective* intelligence significantly exceeds the abilities of individual agents. I expect that further advancements in autonomous LLM agents will make such systems a reality in a few years. Ensuring the safety and alignment of these models is key, and their societal importance calls for interdisciplinary work that combines technical solutions with analysis of broader impact of these models [NeurIPS'24c WS].

◇ **Evaluation of risks and capabilities of frontier models.** We need to have a clear picture of the risks and capabil-

ities of the latest frontier models that are continuously evolving. This is a complex task that involves addressing challenges such as (1) the need for broad coverage of diverse use cases that are evolving over time, (2) the risk of data contamination, since frontier models are trained on nearly all publicly available data, and (3) the difficulty of collecting high-quality data, especially in specialized domains. I believe my past experience as a first author on prominent robustness and alignment benchmarks [NeurIPS'21], [NeurIPS'24a], [arXiv'24:B], will be instrumental for advancing this research direction. Moreover, measuring *risks* inherently assumes a worst-case evaluation which usually requires solving an optimization problem within some *threat model* that defines its constraints. My goal is to formulate clear, well-motivated, task-dependent threat models and develop efficient optimization techniques for both white-box and black-box attacks for text, vision, and voice frontier models. Moreover, I expect that methods that analyze LLM representations can be particularly important for uncovering potentially hidden capabilities and important properties that are not apparent from generated outputs. I also plan to focus on exploring the robustness and safety of specialized reasoning models, such as the `o1` family from OpenAI trained with reinforcement learning to produce detailed reasoning steps.

◇ **Improving reliability of frontier models.** Improving robustness of frontier models to adversarial inputs can be achieved via *fine-tuning* with approaches such as adversarial training and representation-based steering. Importantly, fine-tuning is still feasible to do with academic resources, unlike pretraining of frontier models from scratch. Adversarial training, which has been successful in improving robustness of image classifiers in a computationally efficient way [NeurIPS'20], presents a promising approach for LLMs and multimodal models. However, adapting this framework to text inputs requires developing efficient algorithms for generating *discrete* adversarial inputs during training. I propose to develop novel fine-tuning algorithms for adversarial training in LLMs and multimodal models, focusing on two key areas. First, I aim to design a process tailored for aligned LLMs to mitigate vulnerabilities to prefix and suffix attacks, without substantial decreases in the model capability. Second, I plan to enhance the robustness of text encoders used in multimodal models. These approaches can be further enhanced by representation steering and editing that have proven to be a valuable method for improving robustness and alignment [NeurIPS'24b]. Progress in this direction will contribute to creating more reliable and robust generative models across different data modalities.

◇ **Concluding remarks.** Ensuring responsible development of advanced AI models is extremely important, particularly as they are becoming more autonomous and agentic. Without proper safety measures in place, the potential for harm could outweigh the benefits of this technology. **Thus, it is very important that both capabilities and safety advance at the same rate.** I look forward to establishing a research group to solve hard problems in this space. I believe that my experience in working on adversarial robustness, designing comprehensive evaluations, and understanding complex phenomena through simple models and controlled experiments will be highly relevant for making progress on this research agenda.

References

[NeurIPS'23a] **Maksym Andriushchenko**, Dara Bahri, Hossein Mobahi, and Nicolas Flammarion. "Sharpness-Aware Minimization Leads to Low-Rank Features". *NeurIPS*. 2023.

[ICML'24 WS] **Maksym Andriushchenko**, Francesco Croce, and Nicolas Flammarion. "Jailbreaking Leading Safety-Aligned LLMs with Simple Adaptive Attacks". *Under submission (a short version appeared at the ICML 2024 Workshop on the Next Generation of AI Safety)* (2024).

[ECCV'20] **Maksym Andriushchenko**, Francesco Croce, Nicolas Flammarion, and Matthias Hein. "Square Attack: A Query-Efficient Black-Box Adversarial Attack via Random Search". *ECCV*. 2020.

[ICML'23b] **Maksym Andriushchenko**, Francesco Croce, Maximilian Müller, Matthias Hein, and Nicolas Flammarion. "A Modern Look at the Relationship Between Sharpness and Generalization". *ICML*. 2023.

[NeurIPS'24b WS] **Maksym Andriushchenko** and Nicolas Flammarion. "Does Refusal Training in LLMs Generalize to the Past Tense?" *NeurIPS Workshop on Safe Generative AI Workshop (Oral)* (2024).

[ICML'22] **Maksym Andriushchenko** and Nicolas Flammarion. "Towards Understanding Sharpness-Aware Minimization". *ICML*. 2022.

[NeurIPS'20] **Maksym Andriushchenko** and Nicolas Flammarion. "Understanding and Improving Fast Adversarial Training". *NeurIPS*. 2020.

[NeurIPS'19] **Maksym Andriushchenko** and Matthias Hein. "Provably Robust Boosted Decision Stumps and Trees Against Adversarial Attacks". *NeurIPS*. 2019.

[CVPR'22 WS] **Maksym Andriushchenko**, Xiaoyang Rebecca Li, Geoffrey Oxholm, Thomas Gittings, Tu Bui, Nicolas Flammarion, and John Collomosse. "ARIA: Adversarially Robust Image Attribution for Content Provenance". *CVPR 2022 Workshop on Media Forensics*. 2022.

[arXiv'24:B] **Maksym Andriushchenko**, Alexandra Souly, Mateusz Dziemian, Derek Duenas, Maxwell Lin, Justin Wang, Dan Hendrycks, Andy Zou, Zico Kolter, Matt Fredrikson, Eric Winsor, Jerome Wynne, Yarin Gal, and Xander Davies. "AgentHarm: A Benchmark for Measuring Harmfulness of LLM Agents". *Under submission (arXiv preprint arXiv:2410.09024)* (2024).

[ICML'23a] **Maksym Andriushchenko**, Aditya Vardhan Varre, Loucas Pillaud-Vivien, and Nicolas Flammarion. "SGD with Large Step Sizes Learns Sparse Features". *ICML*. 2023.

[NeurIPS'24a] Patrick Chao*, Edoardo Debenedetti*, Alexander Robey*, **Maksym Andriushchenko***, Francesco Croce, Vikash Sehwag, Edgar Dobriban, Nicolas Flammarion, George J. Pappas, Florian Tramèr, Hamed Hassani, and Eric Wong. "JailbreakBench: An Open Robustness Benchmark for Jailbreaking Large Language Models". *NeurIPS Datasets and Benchmarks Track*. 2024.

[AISTATS'19] Francesco Croce*, **Maksym Andriushchenko***, and Matthias Hein. "Provable Robustness of ReLU networks via Maximization of Linear Regions". *AISTATS*. 2019.

[NeurIPS'21] Francesco Croce*, **Maksym Andriushchenko***, Vikash Sehwag*, Edoardo Debenedetti*, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. "RobustBench: A Standardized Adversarial Robustness Benchmark". *NeurIPS Datasets and Benchmarks Track*. 2021.

[AAAI'22] Francesco Croce, **Maksym Andriushchenko**, Naman D Singh, Nicolas Flammarion, and Matthias Hein. "Sparse-RS: A Versatile Framework for Query-Efficient Sparse Black-Box Adversarial Attacks". *AAAI*. 2022.

[NeurIPS'24c] Francesco D'Angelo*, **Maksym Andriushchenko***, Aditya Varre, and Nicolas Flammarion. "Why Do We Need Weight Decay in Modern Deep Learning?" *NeurIPS*. 2024.

[NeurIPS'24c WS] Joshua Freeman, Chloe Rippe, Edoardo Debenedetti, and **Maksym Andriushchenko**. "Exploring Memorization and Copyright Violation in Frontier LLMs: A Study of the New York Times v. OpenAI 2023 lawsuit". *Under submission (a short version appeared at the NeurIPS 2024 Safe Generative AI Workshop)* (2024).

[NeurIPS'17] Matthias Hein and **Maksym Andriushchenko**. "Formal Guarantees on the Robustness of a Classifier Against Adversarial Manipulation". *NeurIPS*. 2017.

[CVPR'19] Matthias Hein, **Maksym Andriushchenko**, and Julian Bitterwolf. "Why ReLU networks yield high-confidence predictions far away from the training data and how to mitigate the problem". *CVPR*. 2019.

[UAI'22] Klim Kireev*, **Maksym Andriushchenko***, and Nicolas Flammarion. "On the Effectiveness of Adversarial Training Against Common Corruptions". *UAI*. 2022.

[NeurIPS'24a WS] Hao Zhao, **Maksym Andriushchenko**, Francesco Croce, and Nicolas Flammarion. "Is In-Context Learning Sufficient for Instruction Following in LLMs?" *Under submission (a short version appeared at the NeurIPS 2024 Workshop on Adaptive Foundation Models)* (2024).

[ICML'24] Hao Zhao, **Maksym Andriushchenko**, Francesco Croce, and Nicolas Flammarion. "Long Is More for Alignment: A Simple but Tough-to-Beat Baseline for Instruction Fine-Tuning". *ICML*. 2024.

[NeurIPS'24b] Andy Zou, Long Phan, Justin Wang, Derek Duenas, Maxwell Lin, **Maksym Andriushchenko**, Rowan Wang, Zico Kolter, Matt Fredrikson, and Dan Hendrycks. "Improving Alignment and Robustness with Circuit Breakers". *NeurIPS*. 2024.