Maksym Andriushchenko Research Statement

My research focuses on understanding **robustness** and **generalization** in AI models to improve their reliability. As AI models become increasingly capable of acting autonomously and operating with less human supervision, it is essential to ensure that their capabilities and safety advance at the same rate. Towards this goal, **my current research centers on improving the safety, alignment, and reliability of Large Language Models (LLMs)**.

Despite their exceptional capabilities, even state-of-the-art LLMs often produce factually incorrect outputs, make basic reasoning mistakes, and generally lack robustness to input reformulations. Moreover, **LLMs remain highly susceptible to adversarial inputs that can bypass their safety guardrails**, potentially enabling large-scale misuse through automated cyberattacks, targeted misinformation campaigns, and very realistic AI-generated content. The combination of these vulnerabilities and the challenges in controlling LLM outputs poses substantial societal risks, particularly as AI adoption accelerates and AI models operate with greater autonomy. 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 specialized on complex reasoning tasks, autonomous LLM-based agents equipped with external tools, and multi-agent systems.

One key reason for the existing vulnerabilities in LLMs is the difficulty in accurately evaluating their adversarial robustness, as this requires solving an intractable optimization problem over a large search space. My past research shows that **we can transfer many lessons learned from robustness evaluation of small vision models** [ECCV'20], [AAAI'22] **to the LLM setting** [ICLR'25a], [NeurIPS'24a]. Moreover, my work on certified defenses against adversarial examples that come with provable robustness guarantees [NeurIPS'17], [AISTATS'19], [NeurIPS'19] could also suggest methods for making LLM more robust. More generally, evaluating the risks and capabilities of advanced AI models presents significant challenges, including assessing diverse use cases, preventing contamination between training and test data, and organizing 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 [ICLR'25b]—will be instrumental for advancing this research direction.

The existence of adversarial inputs in general can be viewed as a generalization problem: achieving robustness requires ensuring generalization beyond the examples in the training set. My conceptual approach in this direction involves using simple models and controlled experiments to develop a better understanding of generalization mechanisms. 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 on the effect of explicit and implicit regularization for selecting better minima [ICML'22], [NeurIPS'23a], [NeurIPS'24c]. These works share a common theme: **I derive theoretical insights from simplified models and then validate them empirically on complex deep networks**. 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].

Throughout my doctorate, **I worked on AI safety with leading organizations in the field**, including OpenAI, Anthropic, the UK AI Safety Institute, Center for AI Safety, and Gray Swan AI. This resulted in both non-public evaluations that influenced the deployment and risk assessment of models and services at both OpenAI and Anthropic, as well as open academic papers such as Circuit Breakers [NeurIPS'24b] and AgentHarm [ICLR'25b]. My PhD research, which led to these collaborations, was recognized through PhD fellowships from Google and Open Philanthropy, as well as the Patrick Denantes Prize for the best PhD thesis in EPFL's IC department.

In my future work, **I aim to advance the 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. It is likely that the most recent advances in using synthetic data to improve reasoning and planning—such as in the OpenAI `o1` models—will lead to highly capable LLM agents, making this research agenda particularly timely. In what follows, I describe in detail my previous work and then discuss my future research plans to make concrete progress on next-generation AI models' safety and alignment.

Evaluating and Improving Robustness in Deep Learning

One of my main research direction is evaluating and improving adversarial robustness, with an emphasis on designing benchmarks that set the correct targets and enable us to systematically track progress in the field. This line of work also informs the development of better defenses against adversarial inputs.

◇ **Robustness evaluation.** Many defenses against adversarial examples proposed in the literature systematically overestimate their adversarial robustness. To improve upon this, I have made substantial progress toward *provable* adversarial robustness. In [NeurIPS'17] and [AISTATS'19], we proposed one of the first provable guarantees on worst-case robustness for ℓ_p-norm bounded adversarial examples 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 merely 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 [ICLR'25a], we used an approach based on random search, similar to the Square Attack, as a key component of our attacks on frontier LLMs. We pointed out the crucial role of *adaptive* attacks that are designed for a particular model. As a result, we were able to "jailbreak" all frontier LLMs with a 100% attack success rate. This demonstrates that achieving robust safety alignment is highly non-trivial even for the latest LLMs, such as GPT-4o and Claude 3.5 Sonnet.

◇ **Robustness benchmarks.** Well-designed benchmarks are crucial for making concrete, measurable progress, especially in a field like adversarial robustness, where worst-case performance is often significantly overestimated. Together with my collaborators, 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` [ICLR'25b] showed that LLMs trained to refuse harmful user instructions in chat settings do not necessarily maintain their safety alignment when deployed as *agents* with access to external tools (see Figure 1), even against *very basic* adversarial attacks. This misalignment poses significant risks, as such tool-using agents can cause harm not only in the digital space but also in the physical world, for example when a multimodal LLM controls a robot or other physical system.

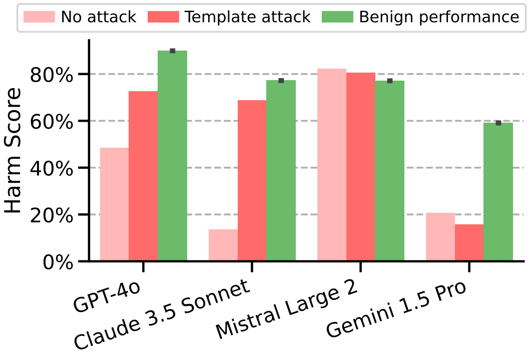

Figure 1: Our most recent work [ICLR'25b] illustrates that even frontier LLMs like GPT-4 and Claude 3.5 Sonnet lack robust safeguards against misuse. Higher scores in **red** imply more harm and scores in **green** serve as a baseline.

◇ **Improving robustness.** Research into stronger adversarial attacks creates the need for more robust defenses against them. In [NeurIPS'20], we developed a better understanding of how and why efficient adversarial training can fail—a phenomenon known as *catastrophic overfitting*. We 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 developed significantly more robust *content authenticity models* [CVPR'22 WS]—a critical 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 showed that it 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

Improving adversarial robustness requires understanding the generalization mechanisms in deep networks, which are often complex because they depend on subtle choices of hyperparameters and architectural components.

◇ **Understanding explicit and implicit regularization.** Generalization in deep learning is influenced not only by explicit regularization like weight decay but also by the choice of optimization parameters, such as batch size or learning rate. The interplay between optimization, training dynamics, and generalization was poorly understood, and it remained unclear why some methods used in practice worked better than others. To disentangle the effect of stochastic noise on generalization, we took a closer look at the implicit regularization mechanism of stochastic gradient descent (SGD) with large learning rates. As our work [ICML'23a] showed, in this regime, SGD induces noisy training dynamics, which drives iterates to better-generalizing minima and makes neural networks learn sparse, input-dependent features. We used diagonal linear networks as simple, theoretically tractable models to explain the training dynamics and reveal the implicit sparsification mechanism present in deep nonlinear networks. An important practical insight is that the stochastic noise—which is proportional to the training loss—should not diminish too quickly, which can be achieved through careful selection of the learning rate schedule. In [ICML'22] and [NeurIPS'23a], we provided a better understanding of Sharpness-Aware Minimization (SAM), an optimization algorithm that biases training towards *flat* minima that have low curvature, and analyzed its effect on the features learned by models. 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 *do not* necessarily generalize better, contrary to a popular intuition in the field. In our most recent work in this direction [NeurIPS'24c], we revisited the role of weight decay in modern deep learning practice, examining its effect on generalization, optimization speed, and training stability with limited floating-point precision. Through this series of works, my coauthors and I used theoretical and empirical tools to develop a systematic understanding of how the choice of optimization hyperparameters crucially affects generalization and the features learned by deep networks.

◇ **Understanding generalization in LLMs.** The *superficial alignment hypothesis* suggests that very few samples are sufficient to teach base LLMs to follow natural language instructions. In our work [ICML'24], we found a very simple recipe for instruction fine-tuning that agrees with this hypothesis: simply fine-tuning on long conversations—which contain more information for learning—outperforms much more complex data selection methods. In [NeurIPS'24a WS], we performed a systematic comparison between two fundamentally different learning approaches: in-context learning and supervised fine-tuning (see Figure 2), including an analysis of their scaling laws. For this project, I also participated in the OpenAI Researcher Access Program where we received access to the *base* GPT-4 model, i.e., the model before instruction fine-tuning or RLHF, which we evaluated in our work. Recently, in [ICLR'25c], we found that simply reformulating harmful requests in the *past tense* is often sufficient to bypass safety measures in frontier LLMs. This can be seen as a clear generalization failure of standard refusal training techniques, such as supervised fine-tuning and RLHF. Finally, our recent work on understanding memorization and its implications for copyright violations in the latest LLMs [NeurIPS'24c WS] also links research on generalization with my research agenda on responsible AI.

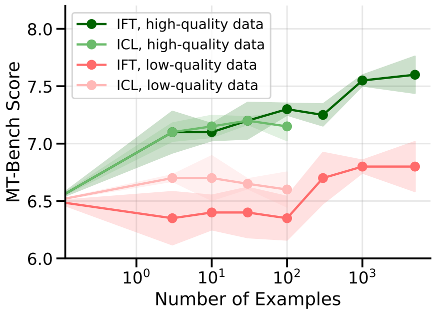

Figure 2: Our work [NeurIPS'24a WS] studies the difference between instruction fine-tuning (IFT) and in-context learning (ICL) for *base* models. The MT-Bench Score measures the generation quality using GPT-4 as a semantic judge.

Future Research Agenda

My future research goal is to advance safety, robustness, and generalization of the next generation of AI models, such as multimodal LLMs, LLMs with enhanced reasoning capabilities, and—most importantly—autonomous LLM agents. The anticipated impact of my future work includes new agent safety benchmarks, novel alignment methods, robust open-source models, reliable agent frameworks, and comprehensive risk and capability evaluation toolkits.

◇ **Science of agent safety.** As LLM agents gain the capability to execute complex, open-ended tasks with minimal supervision, standard alignment methods like RLHF become insufficient. This is well illustrated by my recent work on `AgentHarm` [ICLR'25b] that shows that current LLM agents—including agents based on the latest GPT-4o and Claude 3.5 Sonnet models—can execute highly harmful tasks under simple, template-based adversarial attacks. This highlights the urgency of improving safety and alignment of LLM agents, which have to be *contextual* and can often be more subtle than just refusing harmful requests. For example, agents should be able to download a new program or delete a file at a user's request, but downloading malware or deleting the home folder can be highly harmful. This will require developing novel alignment methods that are highly contextual and extend to both single-agent and multi-agent systems. My research program aims to develop new methods for ensuring that LLM agents reliably comply with safety specifications, in some cases with provable guarantees. I plan to conduct research in the following directions:

- Agent safety benchmarks and evaluation methods. Building on my experience creating AgentHarm, I will construct realistic agent environments with open-ended tasks to analyze how harmful actions—due to jailbreaks, prompt injections, and model misalignment—manifest in long execution traces. Importantly, proper safety evaluations would require adaptive adversarial attacks [ICLR'25a] and carefully selected tasks that can expose accidental model misalignment. Having these benchmarks will also allow us to establish empirical laws describing how safety properties evolve with key parameters, such as model scale, test-time compute, and quality of fine-tuning data.
- Ensuring compliance with safety specifications. This will require teaching models to have a deeper understanding of alignment—specifically, how to recognize which actions are harmful or socially unacceptable in different contexts. In particular, it will be important to teach models to *robustly* follow principles, rules, and legal documents. For test-time approaches, I will work on developing formal guarantees to ensure that the agent's actions are constrained within workflows that are safe according to predefined safety specifications. For training-time methods, I will develop techniques targeting internal representations associated with harmfulness and misalignment, leveraging the insights from our work on Circuit Breakers [NeurIPS'24b].
- Human oversight for long-horizon tasks. Since LLM agents are becoming increasingly capable at executing long-horizon tasks, it is crucial to maintain human oversight over their actions. This is challenging since agent execution traces for such tasks become extremely long and thus intractable to inspect manually. Therefore, it is crucial to develop specialized methods to monitor agents and detect their potential safety failures, using both their outputs and internal activations. Such methods will be useful for model debugging, test-time interventions, and post-hoc detection of misuse and misalignment.

This research program will produce: (1) agent safety benchmarks for realistic workflows, (2) empirical laws governing safety, (3) robustness guarantees for agents, and (4) practical monitoring and intervention tools. My most recent papers on capability-based scaling laws for jailbreaking [arXiv'25a], monitoring sequential harmful actions [arXiv'25b], and benchmarking safety of computer use agents [arXiv'25c] can be seen as first steps in the direction of this agenda.

◇ **Continuous evaluation of risks and capabilities of frontier AI models.** The landscape of frontier AI models is rapidly evolving. Beyond agentic capabilities, they can specialize in various domains, such as reasoning, forecasting, or different data modalities. We need to have a comprehensive picture of the risks and capabilities of these models. 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. My past experience as a (co-)first author on prominent robustness and alignment benchmarks [NeurIPS'21], [NeurIPS'24a], [ICLR'25b] 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 a *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 representation-based methods can be particularly important for uncovering latent capabilities and important properties that are not readily apparent from generated outputs alone. My past work on analyzing the features learned by deep networks will help make progress in this direction.

◇ **Concluding remarks.** Ensuring responsible development of advanced AI models is critical, particularly as they are being deployed as autonomous agents. Without proper safety measures in place, the potential for harm could outweigh the benefits of this technology. **Thus, it is essential that both model capabilities and safety advance at the same rate.** I look forward to establishing a research group that focuses on developing technical solutions to improve AI safety and generalization. I am also interested in exploring the broader implications of AI alignment and pursuing interdisciplinary work in this direction. I plan to continue external collaborations with organizations outside academia, such as frontier LLM labs, non-profit AI safety organizations, and government institutions like the UK AI Safety Institute. 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.

[ICLR'25a] **Maksym Andriushchenko**, Francesco Croce, and Nicolas Flammarion. "Jailbreaking Leading Safety-Aligned LLMs with Simple Adaptive Attacks". *ICLR* (2025).

[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.

[ICLR'25c] **Maksym Andriushchenko** and Nicolas Flammarion. "Does Refusal Training in LLMs Generalize to the Past Tense?" *ICLR* (2025).

[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.

[ICLR'25b] **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". *ICLR* (2025).

[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.

[arXiv'25b] Yueh-Han Chen, Nitish Joshi, Yulin Chen, Maksym Andriushchenko, Rico Angell, and He He. "Monitoring Decomposition Attacks in LLMs with Lightweight Sequential Monitors". *arXiv preprint arXiv:2506.10949* (2025).

[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 2021 Datasets and Benchmarks Track (a short version received a best paper honorable mention at the ICLR 2021 Workshop on Security and Safety in ML Systems).*

[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.

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

[arXiv'25c] Thomas Kuntz, Agatha Duzan, Hao Zhao, Francesco Croce, Zico Kolter, Nicolas Flammarion, and Maksym Andriushchenko. "OS-Harm: A Benchmark for Measuring Safety of Computer Use Agents". *arXiv preprint arXiv:2506.14866* (2025).

[arXiv'25a] Alexander Panfilov, Paul Kassianik, Maksym Andriushchenko, and Jonas Geiping. "Capability-Based Scaling Laws for LLM Red-Teaming". *arXiv preprint arXiv:2505.20162* (2025).

[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.