Maksym Andriushchenko **Teaching Statement**

Ensuring alignment of AI with societal interests is a key challenge of our time. I believe education can play a vital role in achieving this goal—**teaching technical subjects in AI must be combined with promoting active thinking of AI's broader impact**. Moreover, it is important to prepare students for real-world challenges by **adapting the educational process to account for the availability of generative AI**. Previously, I have helped teach nine courses in multiple institutions including *Machine Learning*, *Probability and Statistics*, and *Advanced Algorithms*. Additionally, I have supervised thirteen students at bachelor's, master's, and PhD level, whose work has led to academic distinctions and multiple top-tier publications [NeurIPS'21], [NeurIPS'24c], [ICML'24]. Looking ahead, I am eager to develop and **teach courses spanning responsible AI, AI safety, and LLM agents**, as well as more general courses related to LLMs, machine learning, introductory mathematics, and computer science.

Teaching Philosophy

Rapid advances in generative AI and its wide availability are challenging established educational practices. I believe we need to adapt courses and design assignments to take this into account.

◇ **Emphasizing a data-centric view on AI.** Given the growing importance of AI technology, I would actively encourage students to think critically about its long-term effects on society and explore ways to improve its outcomes by combining both algorithmic *and* data-driven approaches. The challenges of *AI alignment*—determining which values we should optimize for and how—become evident when examining various biases in training data. These range from discriminatory stereotypes to very subtle biases in human preferences that can lead to *sycophancy*, i.e., when a model learns to simply tell users what they want to hear instead of providing truthful answers. Ensuring responsible development of AI systems thus requires examining large amounts of data at all stages of AI development, from collecting pretraining data and human preferences to generating synthetic training data and conducting post-deployment monitoring. The way we provide input-output specifications for these models via training data is just as important as the ML algorithms themselves. Thus, in my courses, I intend to emphasize a more data-centric view of AI development that includes discussions of key data-related topics such as dataset curation, distribution shifts, out-of-distribution generalization, and principles of benchmarking.

◇ **Teaching students to build *with* LLMs.** While getting *a* solution from an LLM becomes easier with each new model release, verifying the correctness of these solutions will remain necessary due to their hallucinations and inaccuracies. Thus, it is still important to teach students the basics and ensure that students are able to implement key primitives, such as backpropagation, from scratch. However, we should also train students to work effectively *with* LLMs. In most courses, there should be assignments for which it should not matter whether solutions are partially AI-generated or not, as long as students can understand and verify their correctness. This would better reflect the real-world conditions, where generative AI is almost always available to help with any task. Thus, I believe we should adapt some assignments to ensure that they cannot be completely solved by LLMs, while still allowing LLMs to be used as helpful tools in the learning process. One concrete approach would be to focus more on larger, project-based assignments where LLMs are useful as assistants *for particular tasks* but where they cannot entirely replace students' scientific thinking and high-level design decisions.

◇ **General approach to teaching.** It is essential to ensure that students are both up-to-date with the latest advances in AI and have strong theoretical foundations that allow them to reason from first principles. My own experience of completing a PhD in the Theory of Machine Learning group at EPFL has shown me that solid theoretical background is crucial for developing a holistic view of the field. It is also key for being able to question existing assumptions and practices—an approach that has been central to my work on generalization in deep learning [ICML'22], [ICML'23b], [NeurIPS'24c]. For project-based assignments, I intend to encourage students to prioritize building lean proofs of concept and prototypes, and then continuously iterate on them. This is particularly easy now, with widely available LLM APIs that are simple to use and allow us to solve most traditional AI tasks with a few API calls. To implement these principles effectively, I strive to create an engaging learning environment during my lectures and seminars that accommodates students from diverse backgrounds and helps them feel connected and comfortable debating and asking questions.

Teaching Experience

I have had the opportunity to serve as a teaching assistant at multiple institutions. At Saarland University, I assisted in teaching the *Neural Networks: Implementation and Application* course led by Prof. Dietrich Klakow in 2017. I also assisted with the *Machine Learning* course taught by Prof. Bernt Schiele at the MPI for Informatics in 2018. At EPFL, I served as a teaching assistant for three courses: *Advanced Algorithms* (2020) taught by Prof. Michael Kapralov, which had over 200 students; *Probability and Statistics* (2021, 2022) taught by Prof. Emmanuel Abbé; and *Machine Learning* (2020, 2021, 2022, 2023) taught by Prof. Martin Jaggi and Prof. Nicolas Flammarion. The *Machine Learning* course was among EPFL's largest, with more than 600 enrolled students. My responsibilities included developing lecture materials, creating assignments, and supervising ML for Science projects. In particular, to keep the course material up-to-date, I designed lectures on adversarial robustness and transformers from scratch. Outside academia, I taught AI lectures to children (ages 7-12) displaced by the war in Ukraine who were temporarily residing in Romania. Since I grew up in Ukraine, I felt that I could meaningfully contribute in this way by teaching basic AI concepts and promoting STEM for children. More generally, I believe it is important to begin outreach activities at a very young age to motivate children—including those from underrepresented communities—to study STEM-related subjects.

Teaching Interests

I would be excited to teach courses ranging from the latest advances in AI safety and LLM agents to basic mathematics and computer science for undergraduate students.

- ◇ A course on responsible AI focusing on AI safety and security, adversarial robustness, privacy, science of evaluations, interpretability, and societal implications of AI. I would also be excited to give a seminar in this area, focusing on the most influential recent papers.
- ◇ A course on LLM agents covering LLM basics, teaching LLMs to use external tools, retrieval-augmented generation, reasoning and planning, software engineering agents, multi-agent systems, alignment techniques, and ethical considerations of using autonomous agents.
- ◇ A broad course on LLMs focusing on architectures, tokenization, pretraining, instruction tuning, in-context learning, advanced reasoning, and multimodal extensions of LLMs.
- ◇ Diverse introductory courses in computer science (e.g., machine learning, algorithms, computer vision, information theory) and mathematics (e.g., probability, statistics, analysis).

Advising Experience and Approach

Mentoring the next generation of AI researchers is one of my key goals. I have worked closely with a number of undergraduate (Mehrdad Saberi, Tiberiu Musat), master's (Oriol Barbany, Etienne Bonvin, Edoardo Debenedetti, Jana Vuckovic, Théau Vannier, Hichem Hadhri, Hao Zhao, Joshua Freeman), and PhD students (Klim Kireev, Francesco d'Angelo, Alexander Panfilov) from EPFL, ETH Zürich, and the University of Tübingen. Our joint work has been accepted at top-tier conferences [NeurIPS'21], [NeurIPS'24c], [ICML'24] and workshops [NeurIPS'24a WS], [NeurIPS'24c WS]. Edoardo's work on `RobustBench` [NeurIPS'21] has received academic recognition through a Best Paper Honorable Mention Prize at an ICLR Workshop. He continued with a master's thesis at Princeton with one of our `RobustBench` collaborators and went on to do a PhD program at ETH Zürich with Prof. Florian Tramèr. Hao's work on understanding instruction fine-tuning of LLMs [ICML'24] received a nomination for EPFL Outstanding Master's Thesis and was independently discovered and covered by MIT Technology Review China. I am pleased to have contributed to his decision to pursue a PhD next year.

My advising approach focuses on understanding each student's interests, establishing a clear vision for problems of mutual interest, and ensuring regular progress toward our goals. Rather than having students assist with small parts of larger projects, I prefer them to take leadership roles—typically as first authors on potential publications. While this approach is more challenging, it is also more rewarding in the long term. Based on my experience, students learn significantly more when they participate in every stage of research, from conducting initial experiments to writing the final paper.

References

[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.

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

[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)*.

[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'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.