Anselm Paulus

anselm.paulus@uni-tuebingen.de • a-paulus.github.io • anselmpaulus
AnselmPaulus • a-paulus • Last updated on February 6, 2026

Current Position

Ph.D. Student in Machine Learning, *University of Tuebingen & Max-Planck-Institute for Intelligent Systems*, Tuebingen, Germany (supervised by Georg Martius on differentiable optimization and LLM safety) 2022 – Present

Education

M.Sc. in Computer Science, *University of Tuebingen* (4.00/4.00) 2021 – 2022

B.S. in Computer Science, *University of Tuebingen* (3.73/4.00) 2018 – 2021

B.S. in Physics, *University of Tuebingen* (3.80/4.00) 2015 – 2021

Previous Positions

Research Scientist Intern, *Meta, Fundamental AI Research (FAIR)*, Menlo Park, CA (with Arman Zharmagambetov on game-theoretical perspectives on LLM safety) 2025 – 2026

Research Scientist Intern, *Meta, Fundamental AI Research (FAIR)*, New York City (with Brandon Amos on amortized optimization for LLM safety) 2023 – 2024

Research Assistant, *Max-Planck-Institute for Intelligent Systems*, Tuebingen, Germany (with Georg Martius on differentiable combinatorial optimization) 2016 – 2019

Publications [Google Scholar: 0.9k+ citations]

Selected publications I am a primary author on are highlighted.

2026

1. *Safety Alignment of LMs via Non-cooperative Games*
   Anselm Paulus, Ilia Kulikov, Brandon Amos, Rémi Munos, Ivan Evtimov, Kamalika Chaudhuri, and Arman Zharmagambetov
   Under submission 2026

2025

2. *AdvPrompter: Fast Adaptive Adversarial Prompting for LLMs* [code]
   Anselm Paulus*, Arman Zharmagambetov*, Chuan Guo, Brandon Amos†, and Yuandong Tian†
   ICML 2025

3. *Hard Contacts with Soft Gradients: Refining Differentiable Simulators for Learning and Control*
   Anselm Paulus*, Andreas René Geist*, Pierre Schumacher, Vít Musil, and Georg Martius
   ICLR 2025

2024

4. *LPGD: A General Framework for Backpropagation through Embedded Optimization Layers* [code]
   Anselm Paulus, Georg Martius, and Vít Musil
   ICML 2024

2023

5. *Backpropagation through Combinatorial Algorithms: Identity with Projection Works* [code]
   Anselm Paulus*, Subham Sekhar Sahoo*, Marin Vlastelica, Vít Musil, Volodymyr Kuleshov, and Georg Martius
   ICLR 2023

2021

6. *CombOptNet: Fit the Right NP-Hard Problem by Learning Integer Programming Constraints* [code]
   Anselm Paulus, Michal Rolínek, Vít Musil, Brandon Amos, and Georg Martius
   ICML 2021 (Spotlight, Oral at LMCA NeurIPS 2020 workshop)

2020

7. *Differentiation of Blackbox Combinatorial Solvers* [code]
   Anselm Paulus*, Marin Vlastelica P.*, Vít Musil, Georg Martius, and Michal Rolínek
   ICLR 2020 (Spotlight)

8. *Optimizing Rank-based Metrics with Blackbox Differentiation* [code]
   Michal Rolínek*, Vít Musil*, **Anselm Paulus**, Marin Vlastelica P., Claudio Michaelis, and Georg Martius
   CVPR 2020 (Oral, best paper award nomination)

9. *Deep Graph Matching via Blackbox Differentiation of Combinatorial Solvers* [code]
   Michal Rolínek, Paul Swoboda, Dominik Zietlow, **Anselm Paulus**, Vít Musil, and Georg Martius
   ECCV 2020

Open Source Repositories

0.6k+ GitHub stars across all repositories.

1. facebookresearch/advgame — ★16 — *AdvGame (Adversarial attacks on LLMs)* 2026
2. a-paulus/softtorch — ★3 — *SoftTorch (library for differentiable operations in PyTorch)* 2025
3. a-paulus/softjax — ★0 — *SoftJAX (library for differentiable operations in JAX)* 2025
4. facebookresearch/advprompter — ★162 — *AdvPrompter (Adversarial attacks on LLMs)* 2025
5. martius-lab/diffcp-lpgd — ★2 — *Lagrangian Proximal Gradient Descent (now merged into CVXPY)* 2024
6. martius-lab/solver-differentiation-identity — ★10 — *Blackbox Identity Differentiation* 2023
7. martius-lab/CombOptNet — ★72 — *CombOptNet* 2021
8. martius-lab/blackbox-deep-graph-matching — ★88 — *Deep Graph Matching* 2020
9. martius-lab/blackbox-backprop — ★346 — *Blackbox Differentiation* 2020

Invited Talks

1. *AdvPrompter: Fast Adaptive Adversarial Prompting for LLMs* — Masaryk University, Brno, Czechia 2024
2. *On differentiable combinatorial optimization* — Dagstuhl Seminar: Machine Learning and Logical Reasoning: The New Frontier 2022

Professional Activities

Reviewing

International Conference on Learning Representations (ICLR): 2022, 2023, 2024, 2025
International Conference on Machine Learning (ICML): 2022, 2023, 2024, 2025
Neural Information Processing Systems (NeurIPS): 2021, 2022, 2025

Neural Information Processing Systems (NeurIPS) DiffCoAlg Workshop: 2025

Skills

Programming	C++, Java, Python
Frameworks	JAX, NumPy, Pandas, PyTorch, SciPy, TensorFlow
Toolbox	Linux, emacs, git, tmux, zsh, uv, vscode