Jonathan Geuter

jonathan.geuter@gmx.de | j-geuter.github.io | www.github.com/j-geuter | linkedin.com/in/jonathan-geuter | Jonathan Geuter

Personal Profile

PhD student in Applied Mathematics at Harvard. Passionate about developing efficient ML algorithms with provable guarantees, particularly in the following areas: LLMs (test-time scaling, alignment, discrete diffusion models), generative modeling (flow matching/diffusion models), and optimal transport. Kempner Graduate Fellow at the Kempner Institute for the Study of Natural & Artificial Intelligence.

Education

Harvard University — *Cambridge, US*
PhD in Applied Mathematics — *Sep 2023 - May 2027 (exp.)*
- Supervised by Prof. David Alvarez-Melis, part of Harvard ML Foundations
- Working on optimal transport for machine learning, LLMs (inference-time algorithms and model alignment), and generative models
- Supported by a Kempner Graduate Fellowship

Harvard University — *Cambridge, US*
MSc in Computer Science — *Sep 2023 - May 2025*

MIT — *Cambridge, US*
Cross-Registered Student — *Sep 2023 - May 2025*

Technische Universität Berlin — *Berlin, Germany*
MSc in Mathematics — *Oct 2020 - Oct 2022*

University of California, Berkeley — *Berkeley, US*
UC Education Abroad Program — *Aug 2018 - May 2019*

Technische Universität Berlin — *Berlin, Germany*
BSc in Mathematics — *Oct 2016 - Sep 2020*

Work Experience

Harvard University — *Cambridge, US*
Teaching Assistant for *Computational Science and Engineering Capstone Project Class* — *Jan 2025 - May 2025*

Jina AI — *Berlin, Germany*
Machine Learning Intern — *May 2023 - Aug 2023*
- Trained state-of-the-art open source language embedding models [HuggingFace link]

Zuse Institute Berlin — *Berlin, Germany*
Research Assistant — *May 2021 - Nov 2022*
- Worked on a Julia package for the Frank-Wolfe algorithm and GNNs in the *Laboratory for Interactive Optimization and Learning*

Technische Universität Berlin — *Berlin, Germany*
Teaching Assistant for *Computer-Oriented Mathematics I and II* — *Oct 2017 - Aug 2018; Oct 2019 - Mar 2021*

Technische Universität Berlin — *Berlin, Germany*
Teaching Assistant for *Calculus I* and *Linear Algebra I* — *Aug 2020 - Oct 2020*

University of California, Berkeley — *Berkeley, US*
Research Intern — *Oct 2018 - Dec 2018*

Publications and Projects

Boomerang Distillation Enables Zero-Shot Model Size Interpolation [link]
S. Kangaslahti, N. V. Nayak*, J. Geuter*, M. Fumero, F. Locatello, D. Alvarez-Melis. *In submission*, 2025.

RoBoN: Routed Online Best-of-n for Test-Time Scaling with Multiple LLMs
J. Geuter, G. Kornhardt. *1st Workshop on Foundations of Reasoning in Language Models at NeurIPS 2025*, San Diego, USA.

Guided Speculative Inference for Efficient Test-Time Alignment of LLMs [link]
J. Geuter, Y. Mroueh, D. Alvarez-Melis. ***Spotlight**, 3rd Workshop for Efficient Systems for Foundation Models at ICML 2025*, Vancouver, Canada.

Entropy-Driven Pre-Tokenization for Byte-Pair Encoding [link]
Y. Hu, N. Liang, D. Zhao, J. Geuter, V. Reddy, C. Schmidt, C. Tanner. In *Proceedings of the Tokenization Workshop (TokShop) at ICML 2025*, Vancouver, Canada.

Universal Neural Optimal Transport [link]
J. Geuter, G. Kornhardt, I. Tomasson, V. Laschos. In *Proceedings of the 42nd International Conference on Machine Learning (ICML 2025)*, Vancouver, Canada.

DDEQs: Distributional Deep Equilibrium Models through Wasserstein Gradient Flows [link]
J. Geuter, C. Bonet, A. Korba, D. Alvarez-Melis. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS 2025)*, Phuket, Thailand.

Jina Embeddings: A Novel Set of High-Performance Sentence Embedding Models [link]
M. Günther, L. Milliken, J. Geuter, G. Mastrapas, B. Wang, H. Xiao. In *Proceedings of the 3rd Workshop for Natural Language Processing Open Source Software at EMNLP (NLP-OSS 2023)*, pages 8-18, Singapore. Association for Computational Linguistics.

A Sinkhorn-NN Hybrid Algorithm for Optimal Transport - *Master's Thesis* *Berlin, Germany*
Technische Universität Berlin and Weierstrass Institute *May 2022 - Oct 2022*

FrankWolfe.jl, CINDy *Berlin, Germany*
Zuse Institute Berlin *May 2021 - Nov 2022*
- Contributed to a Julia package of the Frank-Wolfe algorithm, and a Python implementation of the CINDy algorithm

Nonlinear Korn Inequalities - *Bachelor's Thesis* *Berlin, Germany*
Technische Universität Berlin and Humboldt Universität zu Berlin *Jun 2020 - Sep 2020*

Skills

Programming	Python (PyTorch, transformers, vLLM, NumPy, TensorFlow, JAX, scikit-learn, pandas, multiprocessing, distributed training, etc.), Julia
Miscellaneous	Bash, LaTeX, Git

Achievements

2025	**Grant,** 2025 QuantCo ICML Travel Scholarship	*US*
2025	**Grant,** Hudson River Trading ICML Scholar Award	*US*
2025	**PhD Fellowship,** Kempner Institute Graduate Fellowship	*Cambridge, US*
2021	**Scholarship,** *Deutschlandstipendium* for two semesters	*Germany*
2020	**Scholarship,** *Deutschlandstipendium* for two semesters	*Germany*
2018	**Grant,** for two semesters of full-time study granted by the University of California, Berkeley	*Berkeley, US*

Service

Reviewer NeurIPS (2023, 2025), AISTATS (2024), ICML (2025), ICLR (2025)

Interests

Hobbies	Biking, hiking, running, going to the gym, bouldering, tennis, badminton, kite surfing, writing, chess
Politics	Engaged in the German Green Party; co-founder of European Horizons Chapter TU Berlin
Volunteering	Founded a soup kitchen during my time at Berkeley; ran a math club for elementary school students for a few years
Skiing	I love skiing and am a certified skiing instructor in Germany

Languages

German	Native proficiency	
English	Professional proficiency	*TOEFL iBT: 118/120*
French	Basic proficiency	*certified level B1, 4 years in high school*
Mandarin	Basic proficiency	*5 years in high school*
Spanish	Basic proficiency	*two semesters at Berliner Volkshochschulen*
Italian	Basic proficiency	*one semester at UC Berkeley*