

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. I work on efficient methods for training, post-training, and inference of language models across both autoregressive LLMs and diffusion LLMs. Passionate about alignment, knowledge distillation, novel architectures, diffusion models, and optimal transport. Kempner Graduate Fellow at the Kempner Institute for the Study of Natural & Artificial Intelligence.

Education

Harvard University

PhD in Applied Mathematics

- Supervised by Prof. David Alvarez-Melis, part of Harvard [ML Foundations](#)
- Working on efficient methods for LLMs and dLLMs (pretraining, fine-tuning, and inference), and optimal transport for ML
- Supported by a [Kempner Graduate Fellowship](#)

Cambridge, US

Sep 2023 - May 2028

Harvard University

MSc in Computer Science

- En route to the PhD

Cambridge, US

Sep 2023 - May 2025

MIT

Cross-Registered Student

Cambridge, US

Sep 2023 - May 2025

Technische Universität Berlin

MSc in Mathematics

Berlin, Germany

Oct 2020 - Oct 2022

University of California, Berkeley

UC Education Abroad Program

Berkeley, US

Aug 2018 - May 2019

Technische Universität Berlin

BSc in Mathematics

Berlin, Germany

Oct 2016 - Sep 2020

Work Experience

IMC Trading

Quantitative Research Intern (*incoming*)

Sydney, Australia

Nov 2026 - Jan 2027

Institute of Foundation Models

AI Research Intern

Sunnyvale, US

May 2026 - Sep 2026

Harvard University

Teaching Assistant for *Computational Science and Engineering Capstone Project Class*

Cambridge, US

Jan 2025 - May 2025

Jina AI

Machine Learning Intern

- Trained state-of-the-art open source language embedding models [[HuggingFace link](#)]

Berlin, Germany

May 2023 - Aug 2023

Zuse Institute Berlin

Research Assistant

- Worked on a Julia package for the Frank-Wolfe algorithm and GNNs in the [Laboratory for Interactive Optimization and Learning](#)

Berlin, Germany

May 2021 - Nov 2022

Technische Universität Berlin

Teaching Assistant for *Computer-Oriented Mathematics I and II*

Berlin, Germany

Oct 2017 - Aug 2018; Oct 2019 - Mar 2021

Technische Universität Berlin

Teaching Assistant for *Calculus I* and *Linear Algebra I*

Berlin, Germany

Aug 2020 - Oct 2020

University of California, Berkeley

Research Intern

Berkeley, US

Oct 2018 - Dec 2018

Publications

Unlocking Lossless Speedups in LLMs via Discrete Diffusion [[link](#)]

S. S. Sahoo*, L. Chen*, K. Pham*, J. Geuter*, C. Dwivedi, V. Pimpalkhute, Y. Akhauri, A. Moreno, M. Yurochkin, Z. Wang, M. Elhoushi, N. Dey, S. Bergsma, J. Hestness, J. Thickstun, E. Xing, Z. Liu. *arXiv preprint arXiv:2609.04010*.

Thinking at the Right Size: Amortized Distillation Across Post-Trained LLMs [[link](#)]

Y. Zhou, S. Kangaslahti*, J. Geuter*, N. V. Nayak, M. Fumero, F. Locatello, D. Alvarez-Melis. *Findings of the Association for Computational Linguistics: EMNLP 2026*, Budapest, Hungary.

Understanding Layer Patching in Model Size Interpolation [[link](#)]

S. Kangaslahti, J. Geuter*, N. V. Nayak*, M. Fumero*, F. Locatello, D. Alvarez-Melis. *Advances in Neural Information Processing Systems (NeurIPS) 2026*, Sydney, Australia. Previously: *AdaptFM: Resource-Adaptive Foundation Model Inference (ICML) 2026*, Seoul, South Korea.

Stop Training for the Worst: Progressive Unmasking Accelerates Masked Diffusion Training [\[link\]](#)

J. Kim*, J. Geuter*, D. Alvarez-Melis, S. Kakade, S. Chen. *Proceedings of the 43rd International Conference on Machine Learning (ICML) 2026*, Seoul, South Korea. Previously: *Multimodal Intelligence Workshop at ICLR 2026*, Rio de Janeiro, Brazil.

Boomerang Distillation Enables Zero-Shot Model Size Interpolation [\[link\]](#)

S. Kangaslahti, N. V. Nayak*, J. Geuter*, M. Fumero, F. Locatello, D. Alvarez-Melis. *Proceedings of the International Conference on Learning Representations (ICLR) 2026*, Rio de Janeiro, Brazil. Previously: *3rd Workshop on Unifying Representations in Neural 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. *Proceedings of the International Conference on Learning Representations (ICLR) 2026*, Rio de Janeiro, Brazil. Previously: *3rd Workshop for Efficient Systems for Foundation Models at ICML 2025 (Spotlight)*, Vancouver, Canada.

RoBoN: Routed Online Best-of-n for Test-Time Scaling with Multiple LLMs [\[link\]](#)

J. Geuter, G. Kornhardt. *1st Workshop on Foundations of Reasoning in Language Models at NeurIPS 2025*, San Diego, USA.

Entropy-Driven Pre-Tokenization for Byte-Pair Encoding [\[link\]](#)

Y. Hu, F. Liang, D. Zhao, J. Geuter, V. Reddy, C. Schmidt, C. Tanner. *Proceedings of the Tokenization Workshop (TokShop) at ICML 2025*, Vancouver, Canada.

Universal Neural Optimal Transport [\[link\]](#)

J. Geuter, G. Kornhardt, I. Tomasson, V. Laschos. *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. *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. *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.

Projects

Open-source projects I've contributed to have accumulated over 1.7k stars on GitHub.

K2 Horizon

Institute of Foundation Models

Sunnyvale, US

May 2026 - Sep 2026

- Contributed to [K2 Horizon](#), a suite of open-source language models ranging from 0.9B to 375B, the largest fully open-sourced LLM release of this scale

finetuner (1.5k★)

Jina AI

Berlin, Germany

May 2023 - Aug 2023

- Contributed to [finetuner](#), a repository for end-to-end finetuning of embedding models

FrankWolfe.jl (>100★), CINDy

Zuse Institute Berlin

Berlin, Germany

May 2021 - Nov 2022

- Contributed to a [Julia package](#) of the Frank-Wolfe algorithm, and a [Python implementation](#) of the CINDy algorithm

A Sinkhorn-NN Hybrid Algorithm for Optimal Transport - Master's Thesis

Technische Universität Berlin and Weierstrass Institute

Berlin, Germany

May 2022 - Oct 2022

Nonlinear Korn Inequalities - Bachelor's Thesis

Technische Universität Berlin and Humboldt Universität zu Berlin

Berlin, Germany

Jun 2020 - Sep 2020

Skills

Programming

Python (PyTorch, transformers, SGLang, vLLM, nano-vLLM, NumPy, TensorFlow, JAX, scikit-learn, pandas, Lightning, multiprocessing, verl, distributed training, etc.), Julia

Miscellaneous

Bash, $\LaTeX$, Git, Slurm, AWS, LLM evals (lm-eval, Daytona, Harbor, etc.)

Talks

2025 Invited talk at Munich NLP about guided speculative inference

virtual

2025 Oral presentation at ES-FoMo Workshop at ICML 2025 on guided speculative inference

Vancouver, Canada

Achievements

2026	Grant , 2026 QRT ICML Travel Grant	Seoul, South Korea
2026	Selective Program , G Research Spring into Quant Finance	Sicily, Italy
2025	Grant , 2025 IMC NeurIPS Travel Scholarship	San Diego, US
2025	Grant , 2025 QuantCo ICML Travel Scholarship	Vancouver, Canada
2025	Grant , Hudson River Trading ICML Scholar Award	Vancouver, Canada
2025	PhD Fellowship , Kempner Institute Graduate Fellowship	Cambridge, US
2021	Scholarship , <i>Deutschlandstipendium</i> for two semesters	Berlin, Germany
2020	Scholarship , <i>Deutschlandstipendium</i> for two semesters	Berlin, Germany
2018	Scholarship , 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, 2026), ICLR (2025, 2026)

Interests

Hobbies	Biking, hiking, running, working out, 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