

Benjamin Minixhofer

PhD Candidate in Natural Language Processing, University of Cambridge

bm644@cam.ac.uk · [bmin.ai](#) · [Google Scholar](#) · [GitHub](#)

RESEARCH INTERESTS

Model reuse and retrofitting, adaptive tokenization, malleability of language models, efficient task-specialized AI.

EDUCATION

University of Cambridge

PhD, Computation, Cognition and Language · Oct 2023 – present (expected May 2027)

Language Technology Lab, supervised by [Anna Korhonen](#) and [Edoardo M. Ponti](#) (previously Ivan Vulić).

Johannes Kepler University Linz, Austria

BSc, Artificial Intelligence · Oct 2020 – Feb 2023

Average grade 1.13 (1 = best, 4 = worst).

RESEARCH EXPERIENCE

Allen Institute for AI (Ai2), Seattle, USA

Research Associate · Nov 2025 – Aug 2026

Research Intern, hosted by Luca Soldaini and Valentin Hofmann · Jul 2025 – Oct 2025

Worked on byteifying Olmo models (Bolmo) and contributed to training of Olmo Hybrid.

University of Edinburgh, United Kingdom

Visiting Researcher · Feb 2025 – Apr 2025

Google DeepMind, Zürich, Switzerland

Student Researcher · May 2024 – Sep 2024

Cohere, Remote

Intern of Technical Staff · May 2023 – Oct 2023

Huawei Noah's Ark Lab, London, United Kingdom

AI Research Intern · Jul 2022 – Sep 2022 and Jun 2020 – Sep 2020

Johannes Kepler University Linz, Austria

Student Research Assistant · Jan 2022 – May 2022

H2O.ai, Remote

Data Science Intern · Aug 2021 – Oct 2021

PUBLICATIONS

* denotes equal contribution. Citations: 527, h-index: 11 (Google Scholar, October 2026).

Journal and Conference Papers

1. [Retrofitting Language Models to Operate over Bytes](#). **B. Minixhofer**, T. Murray, T. Limisiewicz, A. Korhonen, L. Zettlemoyer, N. A. Smith, E. M. Ponti, L. Soldaini, V. Hofmann. *Nature*, 2026.
2. [Olmo Hybrid: From Theory to Practice and Back](#). W. Merrill, Y. Li, T. Romero, A. Svete, C. Costello, P. Dasigi, D. Groeneveld, D. Heineman, B. Kuehl, N. Lambert, C. Li, K. Lo, S. Malik, D. Matusz, **B. Minixhofer**, J. Morrison, L. Soldaini, F. Timbers, P. Walsh, N. A. Smith, H. Hajishirzi, A. Sabharwal. *COLM*, 2026.
3. [Navigating the Alignment-Calibration Trade-off: A Pareto-Superior Frontier via Model Merging](#). T. Hu, **B. Minixhofer**, N. Collier. *Findings of ACL*, 2026.
4. [Universal Cross-Tokenizer Distillation via Approximate Likelihood Matching](#). **B. Minixhofer**, I. Vulić, E. M. Ponti. *NeurIPS*, 2025.
5. [Retrofitting Large Language Models with Dynamic Tokenization](#). D. Feher, I. Vulić, **B. Minixhofer**. *ACL*, 2025.
6. [Zero-Shot Tokenizer Transfer](#). **B. Minixhofer**, E. M. Ponti, I. Vulić. *NeurIPS*, 2024.
7. [Segment Any Text: A Universal Approach for Robust, Efficient and Adaptable Sentence Segmentation](#). M. Frohmann, I. Sterner, I. Vulić*, **B. Minixhofer***, M. Schedl*. *EMNLP*, 2024.
8. [CompoundPiece: Evaluating and Improving Decomposing Performance of Language Models](#). **B. Minixhofer**, J. Pfeiffer*, I. Vulić*. *EMNLP*, 2023.
9. [Where's the Point? Self-Supervised Multilingual Punctuation-Agnostic Sentence Segmentation](#). **B. Minixhofer**, J. Pfeiffer, I. Vulić. *ACL*, 2023.
10. [HumSet: Dataset of Multilingual Information Extraction and Classification for Humanitarian Crises Response](#). S. Fekih, N. Tamagnone, **B. Minixhofer**, R. Shrestha, X. Contla, E. Oglethorpe, N. Rekabsaz. *Findings of EMNLP*, 2022.
11. [WECHSEL: Effective Initialization of Subword Embeddings for Cross-Lingual Transfer of Monolingual Language Models](#). **B. Minixhofer**, F. Paischer, N. Rekabsaz. *NAACL*, 2022.
12. [Enhancing Transformers with Gradient Boosted Decision Trees for NLI Fine-Tuning](#). **B. Minixhofer**, M. Gritta, I. Iacobacci. *Findings of ACL*, 2021.

Preprints

1. [Tokenization: A Survey for Modern NLP](#). M. Cогnetta, C. Akiki, P. S. Ammanamanchi, C. Arnett, T. Bauwens, P. Chizhov, P. Delobelle, K. Dobler, D. Elliott, O. Goldman, T. Hiraoka, V. Hofmann, E. Jang, J. Kallini, I. Kesen, S. Kumar, S. Land, J. Libovický, T. Limisiewicz, A. Liu, C. Meister, **B. Minixhofer**, S. Moon, T. Pimentel, D. Pohl, S. Salhan, T. Sawada, P. Stepachev, G. Vico, P. Whittington, X. Yang, Y. Pinter. *Preprint*, 2026.

AWARDS, HONOURS AND FUNDING

- [Google TPU Research Cloud](#) compute grants with quota valued at approximately \$5M total, from June 2021 to December 2026.
- Gemma 3 Research Grant, \$5k.
- Google Cloud PhD Research Grant, \$1k.
- Royal Society Grant (PhD funding), University of Cambridge, 2023–present.
- Top Reviewer, NeurIPS 2025.
- Outstanding Reviewer, EMNLP 2024.
- [Kaggle Gold Medal](#), 9th of 943 teams, chaitin Hindi and Tamil Question Answering.
- [Kaggle Gold Medal](#), solo 5th of 2,023 teams, PetFinder.my Adoption Prediction.

INVITED TALKS AND GUEST LECTURES

- **Tokenizer Transfer.** University of Toronto, 2026.
- **Tokenization in Modern LLMs** (guest lecture). University College London, 2026.
- **Tokenization in Modern LLMs** (guest lecture). University of Edinburgh, 2025.
- **The Past, Present, and Future of Tokenization.** NLIP Seminar, University of Cambridge, Nov 2024.
- **The Past, Present, and Future of Tokenization** (guest lecture). University of Göttingen, Nov 2024.
- **Zero-Shot Tokenizer Transfer.** Google DeepMind, 2024.
- **Zero-Shot Tokenizer Transfer.** Mozilla, 2024.

TEACHING AND SUPERVISION

Teaching

- Supervisions (small-group tutorials) for undergraduates in Computational Linguistics at the University of Cambridge in Michaelmas 2025, Lent 2026 and Easter 2026.
- Guest lectures on tokenization in Master's courses at University College London (2026), the University of Edinburgh (2025) and the University of Göttingen (2024).
- Educational [Kaggle notebooks](#), >250,000 views total.

Supervision

- **Lukas Probst**, Master's thesis, Karlsruhe Institute of Technology (ongoing).
- **Hari Krishnan Kesavan**, Master's thesis, JKU Linz (ongoing).
- **Darius Feher**, Master's thesis, University of Cambridge. Led to a first-author paper at ACL 2025.
- **Markus Frohmann**, Master's thesis, JKU Linz. Led to a first-author paper at EMNLP 2024.
- **David Fleischanderl**, Bachelor's thesis, JKU Linz.

SERVICE

Reviewing. NeurIPS 2025 (Top Reviewer); EMNLP 2024 (Outstanding Reviewer); ACL Rolling Review (2024-); ACL 2026 Student Research Workshop, 1st and 2nd Tokenization Workshop.

Research integrity. Discovered cheating by the first-place team in the Kaggle PetFinder.my competition, which led to their disqualification after an investigation by Kaggle. Covered by [VICE](#) and [The Register](#).

OPEN-SOURCE SOFTWARE

Open-source Python and Rust libraries for NLP with about 2.6k GitHub stars and 3 million downloads in total (1.35M on PyPI, 1.68M on crates.io), including:

- [wtpsplit](#): sentence segmentation models from Where's the Point? and Segment Any Text.
- [tokenkit](#): tokenizer transfer and cross-tokenizer distillation.
- [zett](#): Zero-Shot Tokenizer Transfer.
- [WECHSEL](#): cross-lingual transfer of monolingual language models.
- [GerPT2](#): German GPT-2 models.

PATENTS

- [Implementing a Multi-Layer Neural Network Using a Single Physical Layer of Analog Neurons](#). **B. Minixhofer**, B. Puchinger, E. Haselsteiner, F. Maier, G. Promitzer, P. Jantscher. US Patent 12,524,660 B2, granted 2026; also granted in China ([CN 113196233 B](#), 2024).