

Tokenization in Large Language Models

Valentin Hofmann, Julie Kallini, Sachin Kumar, Jindřich Libovický, Tomasz Limisiewicz,
Elizabeth Salesky

<https://tokenization-tutorial.github.io>

Tokenization defines how data is represented as input and output for machine learning systems, including large language models (LLMs). While tokenization has long been a foundational component of NLP systems, it has historically remained a niche research area, often considered solved. The recent proliferation of LLMs has changed this: as these models are deployed across diverse languages and domains, tokenization-related issues have become increasingly apparent—from cross-lingual representation disparities to efficiency bottlenecks. This tutorial will provide a comprehensive overview of tokenization methods in text, starting from classical word-based and morphological approaches, to the current standard of statistical subword tokenizers, and emerging methods on end-to-end learnable tokenizers. We will also discuss the implications of tokenization choices on model performance, fairness, and efficiency, and present best practices for evaluating and designing tokenizers in multilingual and domain-specific settings.

Valentin Hofmann

email: valentin.hofmann@lmu.de

website: <https://valentinhofmann.github.io>

Valentin Hofmann is an assistant professor at LMU Munich and a principal investigator at the Munich Center for Machine Learning. His research lies at the intersection of natural language processing, linguistics, and social science, with a particular focus on tokenization, evaluation, and fairness. He holds a Ph.D. in Linguistics from the University of Oxford and a Ph.D. in Computer Science from LMU Munich. Previously, he was a postdoctoral researcher at the Allen Institute for AI and the University of Washington.

Julie Kallini

email: kallini@stanford.edu

website: <https://juliekallini.com>

Julie Kallini is a Ph.D. student in the Computer Science Department at Stanford University, advised by Christopher Potts and Dan Jurafsky. Her research focuses on language model architectures and efficiency, computational linguistics, cognitive science, and tokenization. She has published first-author papers at ACL, EMNLP, ICLR, and ICML, and her work received the Best Paper Award at ACL 2024. She also has extensive teaching experience and served as the head teaching assistant for Stanford’s NLP course in 2026.

Sachin Kumar

email: kumar.1145@osu.edu

website: <https://www.shocheen.com/>

Sachin Kumar is an assistant professor at the Ohio State University. He received his Ph.D. at Carnegie Mellon University. His research focuses on building user-centric NLP models that work well for all languages, populations, and individuals, covering research areas such as tokenization unfairness, tokenizer adaptation, and dynamic/learnable tokenization. His work has been published at venues such as ACL, EMNLP, NAACL, EACL, ICLR, and NeurIPS. He has previously co-taught tutorials at WWW

2022 and EMNLP 2023, as well as co-organized workshops at EMNLP 2024, ICML 2025, COLM 2025, ACL 2026, and COLM 2026.

Jindřich Libovický

email: libovicky@ufal.mff.cuni.cz

website: <https://ufal.mff.cuni.cz/jindrich-libovsky>

Jindřich Libovický is a Research Associate at the Institute of Formal and Applied Linguistics (ÚFAL), Charles University, Prague, where he leads a research group on multilinguality and cross-lingual fairness. His research spans multilingual language modeling and machine translation, with particular focus on how language models align across languages, how to design better tokenization and subword segmentation methods for multilingual settings, and how model performance varies across languages in ways that affect fairness. He holds a Ph.D. in Computational Linguistics from Charles University and previously worked as a researcher at LMU Munich.

Tomasz Limisiewicz

email: tomlim@cs.washington.edu

website: <https://tomlimi.github.io>

Tomasz Limisiewicz is a Postdoctoral Researcher at Meta (FAIR) and the University of Washington. His research motivation is developing better and fairer (multilingual) language models. Tomasz's recent work explores more efficient data representation by language modeling, and latent tokenization, applicable across languages and further beyond text. His work has been published in ACL, EMNLP, NAACL, and ICLR conferences.

Elizabeth Salesky

email: elizabeth.salesky@gmail.com

website: <https://esalesky.github.io>

Elizabeth Salesky is a Research Scientist at Google DeepMind. She holds a Ph.D. from the Center for Language and Speech Processing at Johns Hopkins University, where her dissertation on robust open-vocabulary processing with pixel-based text representations won the 2026 AMTA Best Thesis Award. Her research focuses on multilingual speech and language processing, with a focus on robust cross-lingual and cross-modal transfer, visually-situated text, and speech translation. She has co-organized several workshops and a speech translation tutorial at EACL 2021.