Watermarking for Large Language Models

Xuandong Zhao **Yu-Xiang Wang** **Lei Li**

As AI-generated text increasingly resembles human-written content, the ability to detect machine-generated text becomes crucial in both the computational linguistics and machine learning communities. In this tutorial, we aim to provide an in-depth exploration of text watermarking, a subfield of linguistic steganography with the goal of embedding a hidden message (the watermark) within a text passage. We will introduce the fundamentals of text watermarking, discuss the main challenges in identifying AI-generated text, and delve into the current watermarking methods, assessing their strengths and weaknesses. Moreover, we will explore other possible applications of text watermarking and discuss future directions for this field. Each section will be supplemented with examples and key takeaways.

Xuandong Zhao, Ph.D. Candidate in Computer Science at the University of California, Santa Barbara.
email: xuandongzhao@cs.ucsb.edu
website: https://xuandongzhao.github.io
He is advised by Professors Yu-Xiang Wang and Lei Li. His research combines machine learning and natural language processing to build ethical and trustworthy generative AI systems. He has published several papers in leading NLP/ML conferences such as ACL, NAACL, EMNLP, ICML, AISTATS, and UAI. In addition to his research, Xuandong has actively participated in the academic community, serving on the Program Committee for ACL/EMNLP 2022/2023, ICML 2023, NeurIPS 2023, and ICLR 2024. His professional experience extends beyond academia, having completed internships at industry giants including Microsoft, Google, and Alibaba. Xuandong is the recipient of the prestigious Chancellor's Fellowship from UC Santa Barbara. He obtained his Bachelor of Science degree in Computer Science from Zhejiang University in 2019.

Yu-Xiang Wang, Faculty Member of the Computer Science Department at UCSB.
email: yuxiangw@cs.ucsb.edu
website: https://sites.cs.ucsb.edu/yuxiangw
Prior to joining UCSB, he was a scientist with Amazon AI in Palo Alto. Even before that he was with the Machine Learning Department at Carnegie Mellon University and had the pleasure of being jointly advised by Stephen Fienberg, Alex Smola, Ryan Tibshirani and Jing Lei. Over the years Yu-Xiang has worked on a diverse set of problems in the broad area of statistical machine learning, e.g., trend filtering, differential privacy, subspace clustering, large-scale learning / optimization, bandits / reinforcement learning, just to name a few. His most recent quests include making differential privacy practical and developing a statistical foundation for off-policy reinforcement learning.

Lei Li, Assistant Professor in the Language Technology Institute at Carnegie Mellon University
email: leili@cs.cmu.edu
website: https://lileicc.github.io
His research interest lies in natural language processing, machine translation, and AI-powered drug discovery. He received his B.S. from Shanghai Jiao Tong University and Ph.D. from Carnegie Mellon University. His dissertation work on fast algorithms for mining co-evolving time series was awarded ACM KDD best dissertation (runner up). His recent work on AI writer Xiaomingbot received 2nd-class award of Wu Wen-tsün AI prize in 2017. He is a recipient of ACL 2021 best paper award, CCF Young Elite award in 2019, and CCF distinguished speaker in 2017. His team won first places for five language translation directions in WMT 2020 and the best in corpus filtering challenge. Previously, he worked at

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 5: Tutorial Abstracts), pages 10–11

ByteDance as the founding director of AI Lab. He has served organizers and area chair/senior PC for multiple conferences including KDD, ACL, EMNLP, ICML, ICLR, NeurIPS, AAAI, IJCAI, and CIKM. He has started ByteDance's machine translation system, VolcTrans and many of his algorithms have been deployed in production.