Synthetic Data in the Era of Large Language Models

Vijay Viswanathan, Xiang Yue, Alisa Liu, Yizhong Wang and Graham Neubig

Website: https://synth-data-acl.github.io/

Progress in natural language processing has historically been driven by better data, and researchers today are increasingly using "synthetic data" - data generated with the assistance of large language models - to make dataset construction faster and cheaper. However, most synthetic data generation approaches are executed in an ad hoc manner and "reinvent the wheel" rather than build on prior foundations. This tutorial seeks to build a shared understanding of recent progress in synthetic data generation from NLP and related fields by grouping and describing major methods, applications, and open problems. Our tutorial will be divided into four main sections. First, we will describe algorithms for producing high-quality synthetic data. Second, we will describe how synthetic data can be used to advance the general-purpose development and study of language models. Third, we will demonstrate how to customize synthetic data generation to support scenario-specific applications. Finally, we will discuss open questions about the production and use of synthetic data that must be answered to overcome some of their current limitations. Our goal is that by unifying recent advances in this emerging research direction, we can build foundations upon which the community can improve the rigor, understanding, and effectiveness of synthetic data moving forward.

Vijay Viswanathan, PhD Student, Carnegie Mellon University
Email: vijayv@andrew.cmu.edu
Website: https://www.cs.cmu.edu/ vijayv
Vijay is a PhD student at Carnegie Mellon University, where he works with Sherry Tongshuang Wu and Graham Neubig. He is interested in making AI models more reliable at following specifications of behavior (e.g. task descriptions or instructions), primarily by using synthetic data to achieve this goal. His research received an Outstanding Demo Paper award at ACL 2022 and he received an NEC Student Research Fellowship in 2022.

Xiang Yue, Postdoctoral Fellow, Carnegie Mellon University
Email: xyue2@andrew.cmu.edu
Website: https://xiangyue9607.github.io/
Xiang Yue is a postdoctoral fellow at Carnegie Mellon University, specializing in natural language processing (NLP). His work focuses on advancing the reasoning capabilities of large language models (LLMs) through a data-centric approach. Xiang's research has earned several awards, including a Best Paper Finalist recognition at CVPR 2024, a Best Paper Honorable Mention at ACL 2023, and a Best Paper Award at IEEE BIBM 2021, all centered around (synthetic) data generation. He was also a recipient of the Carnegie Bosch Postdoctoral Fellowship and was recognized as a rising star at the 2024 UMASS Generative AI Workshop.

Alisa Liu, PhD Student, University of Washington
Email: alisaliu@cs.washington.edu
Website: https://alisawuffles.github.io/
Alisa Liu is a PhD student at the University of Washington, working with Yejin Choi and Noah Smith. Her research interests includes developing algorithms for text generation, particularly as a tool for data creation. She is supported by the NSF Graduate Research Fellowship and OpenAI SuperAlignment Fellowship.

Yizhong Wang, PhD Student, University of Washington
Email: yizhongw@cs.washington.edu

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 5: Tutorial Abstracts), pages 11–12
July 27, 2025 ©2025 Association for Computational Linguistics

Website: https://homes.cs.washington.edu/ yizhongw/

Yizhong Wang is a PhD student at the University of Washington, advised by Hannaneh Hajishirzi and Noah Smith. He is also a student researcher at the Allen Institute for Artificial Intelligence (AI2), working on building open language models. His research focuses on the fundamental data challenges in AI development and algorithms centered around data. He has won multiple paper awards, including ACL 2024 Best Theme Paper, CCL 2020 Best Paper, and ACL 2017 Outstanding Paper. He was the co-organizer of the Student Research Workshop at ACL 2020 and the Instruction Tuning and Instruction Following Workshop at NeurIPS 2023.

Graham Neubig, Associate Professor, Carnegie Mellon University

Email: gneubig@cs.cmu.edu

Website: https://www.phontron.com/

Graham Neubig is an associate professor at the Language Technologies Institute of Carnegie Mellon University. His research focuses on natural language processing, with a particular interest in fundamentals, applications, and understanding of large language models for tasks such as question answering, code generation, and multilingual applications. He has published over 270 papers in *ACL venues, and won 5 best paper or honorable mention awards. He has taught tutorials at several *ACL conferences, such as a tutorial on neural networks at EMNLP, and a tutorial on code generation at NAACL.