The Data Frontier for Large Language Models: Selection, Synthesis, and Tools

Lijun Wu, Wentao Zhang, Conghui He

As the development of Large Language Models (LLMs) matures, the focus of the research community is undergoing a critical shift from a purely model-centric to a data-centric paradigm. It is now evident that the quality, diversity, and composition of training data—not merely its scale—are the primary drivers of a model's advanced capabilities, from complex reasoning to reliable instruction following. However, acquiring and curating such high-quality data remains a significant bottleneck. This tutorial provides a comprehensive and practical guide to the state-of-the-art in data research directions for LLMs. We structure the tutorial around the two core pillars of modern data strategy: intelligent data selection and advanced data synthesis. In the first part, we delve into methods for curating the most valuable information from vast, noisy datasets, covering techniques like LLM-as-a-judge for automated quality filtering and active learning for maximizing annotation efficiency. The second part explores the synthetic data revolution, detailing paradigms that range from generating complex reasoning traces (e.g., Chain-of-Thought) to deploying sophisticated multi-agent workflows that can autonomously create high-quality, diverse instruction data from raw seeds. Finally, we will conclude with a practical overview of open-source tools and platforms that facilitate these data-centric workflows, empowering researchers and practitioners to build better models through better data. Attendees will leave with a principled framework and actionable insights for designing and implementing the advanced data strategies required to build the next generation of powerful, specialized, and aligned LLMs.

Lijun Wu
email: lijun_wu@outlook.com
website: https://apeterswu.github.io/
Dr. Lijun Wu is a Young Scientist in Shanghai AI Laboratory. Previously, he was a Research Scientist in ByteDance, a Senior Researcher in Microsoft Research. He got the Ph.D. degree from Sun Yat-sen University (SYSU), and was a member of joint Ph.D. program between SYSU and MSRA. His research interests are on AI/LLMs (e.g., data-centric intelligence, SFT/RL), AI4Science (e.g., LLM4Science, scientific reasoning).

Wentao Zhang
email: weantao.zhang@pku.edu.cn
website: https://zwt233.github.io/
Dr. Wentao Zhang is an assistant professor (Principal Investigator/PhD Advisor) in the Center of Machine Learning Research at Peking University (PKU), and he leads the Data-centric Artificial Intelligence (DCAI) group. Wentao's research focuses on DCAI, LLM, AI systems and AI4Science. Wentao is the contributor or designer of several system projects, including DataFlow, MinerU, and Angel. Before joining PKU, wen- tao worked as a research fellow with Prof. Jian Tang at Montreal Institute for Learning Algo- rithms (Mila, led by Prof.Yoshua Bengio), and he received his Ph.D. degree in CS at PKU, supervised by Prof. Bin Cui.

Conghui He
email: heconghui@pjlab.org.cn
website: https://conghui.ai/

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 5: Tutorial Abstracts), pages 3–4

Dr. Conghui He is a Young Leading Scientist at the Shanghai AI Lab and an Adjunct Doctoral Supervisor at School Shanghai Jiao Tong University. Recognized as a National-level Young Talent, he hold a Ph.D. from Tsinghua University and was a visiting researcher at Stanford University and Imperial College London. He is the creator of MinerU, the world's leading open-source data engine for large models. Additionally, he oversees a dedicated data team that curates high-quality datasets for leading models such as InternLM and InternVL.