Title: Selecting High-Quality Data for Training Large Language Models

This talk will focus on the importance of high-quality training data for large language models (LLMs), arguably the most crucial and least transparent aspect of recent LLM developments. The talk is divided into three parts:

First, I will discuss the role of high-quality training data at different stages of LLM training, from pre-training to supervised fine-tuning and reinforcement learning with human feedback. I aim to provide a brief overview of state-of-the-art techniques and practices for curating and selecting training data, at least for the "known" knowledge in open-source models. Despite tremendous success, I argue that most common practices are ad-hoc and difficult to understand, and advocate for more principled and systematic approaches to data-centric research in the era of LLMs.

In the second and third parts, I will discuss two research projects from my lab, focusing on the selection of data for pre-training and instruction tuning. I will present QuRating, a simple framework for selecting pre-training data that captures the abstract attributes of texts humans intuitively perceive. We demonstrate that using state-of-the-art LLMs (e.g., GPT-3.5-turbo) can discern these qualities in pairwise judgments and emphasize the importance of balancing quality and diversity. We have created QuRatedPajama, a dataset comprising 260 billion tokens with fine-grained quality ratings, and show that sampling according to these ratings improves perplexity and in-context learning.

Finally, I will introduce LESS, a method that effectively estimates the influence of data for identifying relevant instruction-tuning data points for specific applications (a setting we call "targeted instruction tuning"). LESS is efficient, transferable (allowing for the use of a smaller model in data selection), optimizer-aware (compatible with Adam), and easy to interpret. We demonstrate that training with a LESS-selected 5% of the data often outperforms training with full datasets across a variety of downstream tasks.