FUDAN UNIVERSITY

Applied Math Ph.D. Seminar

An Analysis for Reasoning Bias of Language Models with Small Initialization

Speaker: Junjie Yao (Shanghai Jiao Tong University)

Time: 2025-05-29, 16:10 to 17:00

Location: Rm 1801, Guanghua East Tower

Advisor: Zhiqin Xu (Shanghai Jiao Tong University)

Abstract: Transformer-based Large Language Models (LLMs) have revolutionized Natural Language Processing by demonstrating exceptional performance across diverse tasks. This study investigates the impact of the parameter initialization scale on the training behavior and task preferences of LLMs. We discover that smaller initialization scales encourage models to favor reasoning tasks, whereas larger initialization scales lead to a preference for memorization tasks. We validate this reasoning bias via real datasets and meticulously designed anchor functions. Further analysis of initial training dynamics suggests that specific model components, particularly the embedding space and self-attention mechanisms, play pivotal roles in shaping these learning biases. We provide a theoretical framework from the perspective of model training dynamics to explain these phenomena. Additionally, experiments on real-world language tasks corroborate our theoretical insights. This work enhances our understanding of how initialization strategies influence LLM performance on reasoning tasks and offers valuable guidelines for training models.