

COMPUTER SCIENCE

PhD Dissertation Defense

Towards Robust Large Language Models via Data Synthesis, Selection, and Augmentation

Guangzeng Han

Advisor: Dr. Xiaolei Huang



October 23, 2026 | 1:00–3:00 p.m. Central Time



ABSTRACT

Large language models have achieved remarkable capabilities, but reliable learning remains difficult when training information is imperfect. Teacher supervision may be poorly suited to smaller students, instruction datasets may contain redundant or unhelpful examples, and synthetic data generated with fixed prompts may exhibit limited semantic diversity.

This dissertation investigates how data synthesis, selection, and augmentation can improve learning reliability and generalization. It introduces ranking-based supervision for knowledge distillation, influence-based selection for instruction tuning, and a genetic approach to generating diverse examples through semantic attribute recombination and mutation.

Together, these findings show that designing training information to fit the learner and task can improve the robustness and generalization of language models without simply increasing data volume.



BIO

Guangzeng Han is a PhD candidate in Computer Science at the University of Memphis. His research focuses on natural language processing and large language models, with an emphasis on data-centric approaches to improving model robustness. His work has appeared at premier NLP conferences, including ACL, EMNLP, and NAACL. He also received a Best Paper Award at the IEEE International Conference on Healthcare Informatics (ICHI).



JOIN VIA MICROSOFT TEAMS

Meeting ID: 236 656 445 135 312

Passcode: ja2nY6og