VirginiaTech
VIRGINIA POLYTECHNIC INSTITUTE AND STATE UNIVERSITY

Computer Science Seminar Series

National Capital Region

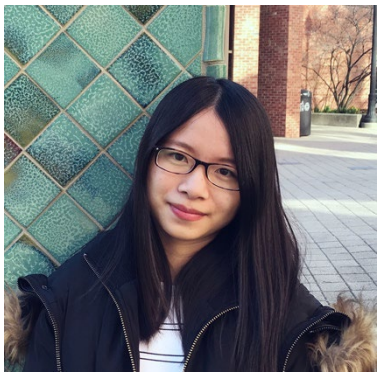
Towards Enhancing the Utilization of Large Language Models for Humans

Speaker: Prof. Ziyu Yao
George Mason University
Friday, March 22, 2024
11:15AM- 12:15PM, NVC 213

Abstract

Large language models, or LLMs, have rapidly transformed the landscape of language technologies. In this talk, I will present two projects in my group towards further enhancing the utilization of LLMs. In the first project, we look into the "cost efficiency" of LLMs and investigate how to save monetary costs for users querying LLM through APIs. Our work proposes a "cascade" of LLMs chaining one weaker LLM with a stronger one, augmented by a mixture-of-thought (MoT) representation to decide when to query the stronger LLM. Our method yields comparable task performance but consumes only 40% of the cost as using only the stronger LLM (GPT-4). In the second project, we turn to the "accessibility" of LLMs and study how to optimize the user prompts to LLMs so as to further democratize the human access to this advanced technique. To this end, we propose an approach that iteratively rewrites human prompts for individual task instances following an innovative manner of "LLM in the loop". Our approach shows to significantly outperform both naive zero-shot approaches and a strong baseline which refines the LLM outputs rather than its input prompts. Finally, I will conclude the talk by presenting other ongoing effort around LLMs in my group.

Biography



Dr. Ziyu Yao (https://ziyuyao.org/) is an Assistant Professor in the Department of Computer Science at George Mason University, where she co-leads the George Mason NLP group (https://nlp.cs.gmu.edu/). Her research covers large language models, semantic parsing/code generation, question answering, and human-NLP/AI interaction, and has been funded by National Science Foundation, Virginia Commonwealth Cyber Initiative, and Microsoft Accelerating Foundation Models Research Program, among others. She has served as an area chair and organized workshops (NLP4Prog, SUKI) at ACL/EMNLP/NAACL. Prior to George Mason, she graduated with Ph.D. in Computer Science and Engineering from the Ohio State University in 2021, where she was awarded the prestigious Presidential Fellowship. She was also selected as a rising star in EECS in 2021.