Department of Computer Science, The University of Hong Kong

CS Seminar

Harmless, Helpful, and Honest LM Agents

Yi Ren FUNG
University of Illinois, Urbana-Champaign

Date:
Dec 4, 2023
Monday
11:00 am -12:30 pm (HKT)

Venue:
Room 308
Chow Yei Ching Building
The University of Hong Kong

Abstract:

In recent years, language models have made significant advancements, achieving high performance on a large variety of tasks, including question answering, summarization, scientific reasoning, procedural understanding, and action planning, and with strong zero-shot/few-shot capability as well, bolstered by model scaling and novel training techniques. Despite these exciting progress, ensuring language models align with fundamental constitutional principles remains a critical challenge. In this talk, we begin by introducing norm discovery as a feasible solution for explainable detection of norm violation occurrences and guiding harmless language model response generations. Then, we broaden our discussion to encompass not only the prioritization of harmlessness but also the exploration of state-of-the-art frameworks aimed at enhancing the utility and helpfulness of language model agents, which includes the development of specialized tools through language model abstract reasoning as well as customized tool retrieval systems. Moreover, we delve into advanced techniques for identifying information inconsistencies in textual or multimedia content, along with extensions to mitigate the phenomenon of hallucination in model prediction especially when handling data that lies beyond the model's parametric knowledge boundary. Our presentation aims to offer an invigorating view of large language models as digital agents that are not only technologically advanced and functionally robust but also sociocultural-aware and ethically grounded, reinforcing their relevance and responsibility in our increasingly interconnected world.

About the Speaker:

Yi R. Fung is a final year PhD student at the University of Illinois Urbana-Champaign, advised by Prof. Heng Ji. Her research centers on language modeling and multimedia knowledge reasoning frameworks that are aligned with human intents, for harmless, helpful, and honest information communication. Yi has previously organized well-recognized tutorial sessions on the frontier of fighting fake news and understanding information with varying shades of truth, at top-tier conferences such as KDD'22. She is also a recipient of the NAACL'21 best demo paper, as well as several prestigious university-level awards. For more information, please feel free to check out her website at https://yrf1.github.io/.

All are welcome!
For enquiries, please call 2859 2180 or email enquiry@cs.hku.hk
Department of Computer Science
The University of Hong Kong