POSTGRADUATE SEMINAR SERIES

Topic Defence Seminar

Topic Title: **Gender Bias Mitigation for Large Language Models applying Retrieval Augmented Generation**

Presenter:

Ms. MA Xiaorui

MPhil student of Computing and Decision Sciences

Abstract :

Large Language Models (LLMs) have become widely used in natural language processing, providing remarkable capabilities for comprehending and generating human-like text. It also makes the study of fairness and ethical considerations in LLMs a growing field of research. The societal biases that LLMs may reinforce are problematic, especially in the context of gender bias, where biased inputs have been revealed to result in prejudiced outputs. The problem is worsened in systems employing Retrieval Augmented Generation (RAG), a technique that supplements a language model's knowledge by retrieving information relevant to the input query from an external database. While RAG improves the model's capabilities, it also risks spreading gender biases found in retrieval sources if the search queries are sensitive or prejudiced. Although efforts have been made to reduce gender bias by changing prompt inputs, the examination of bias introduced during retrieval remains unexplored. This thesis addresses the issue of gender bias in RAG-based systems. For two types of queries prone to producing biased retrieval results, I propose debiasing pipelines built specifically to reduce the bias introduced at the retrieval stage. At the core of this pipeline is a Generative Adversarial Network (GAN) based debiasing generator, which is tailored to reformulate the retrieved content in a manner that both retains the relevant context and removes the gender indication.

Date : 24 April 2024, Wednesday
Time : 16:00 – 18:00
Venue : SEK104, 1/F, Simon & Eleanor Kwok Building
Language : English

*** All are Welcome ***

For enquiries, please contact 2616-8373 or by email (hkibs@ln.edu.hk).

Due date: 25 April 2024