Contribution ID: **57** Type: **not specified**

On Trustworthiness of Large Language Models

Thursday, 16 May 2024 15:45 (1 hour)

Past few years have witnessed significant leaps in capabilities of Large Language Models (LLMs). LLMs of today can perform a variety of tasks such as summarization, information retrieval and even mathematical reasoning with impressive accuracy. What is even more impressive is LLMs'ability to follow natural language instructions without needing dedicated training datasets. However, issues like bias, hallucinations and lack of transparency pose a major impediment to wide adoption of these models. In this talk, I will review how we got from "traditional NLP"to today's LLMs, and some of the reasons behind trustworthiness issues surrounding LLMs. I will then focus on a single issue —hallucinations in factual question answering —and show how artifacts associated with model generations can provide hints that the generation contains a hallucination.

Type of presentation

Invited Talk

Primary author: ZAFAR, Muhammad Bilal (Ruhr University Bochum and Research Center for Trustworthy Data Science and Security)

Presenter: ZAFAR, Muhammad Bilal (Ruhr University Bochum and Research Center for Trustworthy Data Science and Security)

Session Classification: Keynote

Track Classification: Spring Meeting