UGCS/FDS11/E05/25

Title of Project:

(English)	Harnessing Large Language Models for Social Trust: Automated Fact-checking for Public Health Informatics by Large Language Models
(Chinese)	利用大語言模型促進社會信任：透過大語言模型對公共衛生資訊學進行自動事實檢查

In the recent digital environment, the rapid spread of false information on online social media platforms presents a significant challenge. The World Health Organization described the health misinformation during the COVID-19 pandemic as an “infodemic”, highlighting its global scale and impact. At the same time, the emergence of large language models (LLMs) has introduced concerns about hallucination issues, further complicating the problem of digital information inaccuracies. While social media serves as a vital tool for information sharing, it often accelerates the spread of rumors, conspiracy theories, and pseudoscience. The consequences are far-reaching, affecting individuals, society, and nations through financial losses, social unrest, and poor decision-making. The complexity and scale of this problem highlight the inadequacy of our current capabilities to effectively identify misinformation. In response, several organizations and government agencies have taken steps to address the problem. For example, the Semantic Forensics program launched by DARPA in 2021 aims to develop technologies to detect and classify fake media. Similarly, recent initiative of the United States Food and Drug Administration to create a Rumor Control hub represents a proactive approach to countering health-related misinformation. These efforts highlight the need for an interdisciplinary approach in responding to misinformation.

Developing and implementing robust fact-checking mechanisms is critical to preserving the integrity of information and maintaining public trust in an increasingly connected world. Current methods for identifying and addressing misinformation are insufficient to handle its scale and complexity. Human fact-checkers struggle to process the vast volume of content generated on online platforms, and static fact-checking models often fail to adapt to the evolving tactics of misinformation spreaders. Additionally, the subjective nature of fact-checking introduces biases, particularly in cases involving framing or context manipulation. Detecting whether individuals are intentionally disseminating false information to incite others is essential for identifying the sources of misinformation and restoring public trust.

This project focuses on addressing the critical challenge of misinformation in the context of public health informatics. It aims to develop innovative analytical frameworks using LLMs. The first proposed approach targets the verification of truthfulness in health-related content. A generative artificial intelligence framework will be developed, leveraging LLMs with few-shot learning to construct semantic health knowledge graphs and enhance semantic reasoning. We will incorporate graph-based retrieval-augmented generation to address the hallucination issue common in LLMs and the limitations of static training data. By utilizing updated semantic health knowledge graphs, the framework will ensure that fact-checking is informed by the latest and most reliable medical news and health information. The second approach focuses on analyzing intentionally manipulated or fabricated content. Multi-agent LLMs will be employed to perform sentiment analysis, allowing for the identification of untrustworthy sources. This method will assess the intent behind misinformation and evaluate its potential impact on public health discourse. Our project seeks to provide robust solutions for automated fact-checking in the health domain. The outcomes of this research will contribute to public health informatics by improving the reliability of health information, mitigating the spread of false information, and fostering social trust in digital communication systems.