Evaluating the reliability of ChatGPT answers for safety management in chemical handling facilities

Dong Geon Lee [a], Sejong Bae [b] and Jong Bae Baek [a]

[a] Department of Safety Engineering, Korea National University of Transportation, Chungbuk, Korea
[b] Department of Medicine, University of Alabama at Birmingham, United States
Email: jbbaek@ut.ac.kr

Abstract: In recent years, the rapid development of Natural Language Processing (NLP) and Artificial Intelligence (AI) has led to the widespread adoption of NLP models such as Chat-Generative Pre-trained Transformer (ChatGPT) across various digital platforms. However, the dependability of ChatGPT's responses has not been evaluated for utilization in handling Hazardous chemicals for safety management purposes.

To evaluate the reliability of the answers, we have asked ChatGPT with content from the safety management certification exam administered by the Board of Certified Safety Professionals (https://www.bcsp.org/). This provides an initial evidence-based responses provided by ChatGPT for the reliability of the responses.

We presented ChatGPT with questions from the safety management certification exam, gathered responses, and compared them with the correct answers. The difference between ChatGPT's exam score and the passing grade were compared.

By evaluating the reliability of ChatGPT's answers, we have determined whether it can be used for providing information for safety management in facilities handling hazardous chemicals in the future. Current ChatGPT version is 3.5. Further research will be needed to assess the reliability of the upcoming version 4.0 and other NLP models.

ACKNOWLEDGEMENTS

This work was supported by the Graduate School of Chemical Safety Management Specialization, funded by the Ministry of Environment.

REFERENCES

OpenAI, GPT-4 Technical Report, 2023.
Tiffany H. Kung, et. al., 2023, Performance of ChatGPT on USMLE: Potential for AI-Assisted Medical Education Using Large Language Models, PLOS Digital Health.

Keywords: *ChatGPT, artificial intelligence, reliability*