Expert Evaluation of ChatGPT Performance for Risk Management Process based on ISO 31000 Standard

M.K.S. Al-Mhdawi[1]; Abroon Qazi[2]; Ammar Alzarrad[3]; Nicholas Dacre[4]; Farzad Pour Rahimian[5]; Mohanad K. Buniya[6] and Hanqin Zhang[7]

[1]Lecturer, School of Computing, Engineering & Digital Technologies, Teesside University, Middlesbrough TS1 3BX, UK; Researcher, Department of Civil, Structural and Environmental Engineering, Trinity College Dublin, The University of Dublin, Dublin D02 PN40, Ireland. ORCID: https://orcid.org/0000-0001-5870-0323. E-mail: m.al-mhdawi@tees.ac.uk (corresponding author)
[2]Associate Professor, School of Business Administration, American University of Sharjah, UAE. E-mail: aqazi@aus.edu
[3]Assistant Professor, Department of Civil Engineering, Marshall University, USA. E-mail: alzarrad@marshall.edu
[4]Associate Professor, Southampton Business School, University of Southampton, UK. E-mail: Nicholas.Dacre@southampton.ac.uk
[5]Professor, School of Computing, Engineering & Digital Technologies, Teesside University, Middlesbrough TS1 3BX, UK. Email: F.Rahimian@tees.ac.uk
[6]Researcher**,** Department of Civil and Environmental Engineering, University Technology Petronas, Malaysia. E-mail: mohanad_18000491@utp.edu.my
[7]PhD student, Southampton Business school, University of Southampton, UK. E-mail: hz2u20@soton.ac.uk

Abstract

ChatGPT is widely known for its ability to facilitate knowledge exchange, support research endeavours, and enhance problem-solving across various scientific disciplines. However, to date, no empirical research has been undertaken to evaluate ChatGPT's performance against established standards or professional guidelines. Consequently, the present study aims to evaluate the performance of ChatGPT for the risk management (RM) process based on ISO 31000 standard using expert evaluation. The authors (1) identified the key indicators for measuring the performance of ChatGPT in managing construction risks based on ISO 31000 and determined the key assessment criteria for evaluating the identified indicators using a focus group session with Iraqi experts; and (2) quantitatively analysed the level of performance of ChatGPT under a fuzzy environment. The findings indicated that ChatGPT's overall performance was high. Specifically, its ability to provide relevant risk mitigation strategies was identified as its strongest aspect. However, the research also revealed that ChatGPT's consistency in risk assessment and prioritization was the least effective aspect. This research serves as a foundation for future studies and developments in the field of AI-driven risk management, advancing our theoretical understanding of the application of AI models like ChatGPT in real-world risk scenarios.

Keywords: ChatGPT, ChatGPT Performance, AI, Risk, Risk management, ISO 31000

Cite this Paper
Al-Mhdawi, M.K.S., Qazi, A., Alzarrad, A., Dacre, N., Rahimian, F.P., Buniya, M.K. and Zhang, H., 2023. Expert Evaluation of ChatGPT Performance for Risk Management Process based on ISO 31000 Standard. The 33rd European Safety and Reliability Conference, Southampton, UK, 3-8 September 2023.

1. Introduction

ISO 31000 is a widely recognised standard for risk management (RM) that provides a comprehensive framework for identifying, analysing, evaluating, treating, monitoring, and communicating risks (Lalonde and Boiral, 2012). Depending on the adopted RM practices and their effectiveness, ISO 31000 standard may be approached differently. Some RM personnel may choose to focus on specific components of the risk management process, such as risk analysis or risk communication. In contrast, others may adopt a more comprehensive approach that considers the entire process, from planning to monitoring and review. One tool that can provide deeper insights into each step of the ISO 31000 process is language models, such as Chat Generative Pre-training Transformer (GPT) known as ChatGPT. It can analyse large volumes of text data, including RM plans, reports, and other relevant documents, to identify patterns and

The research presented in this pre-print has been submitted to the
33rd European Safety and Reliability Conference (ESREL 2023)

trends that may not be immediately apparent to human experts (White et al. 2023; Nikolic et al. 2023). For example, ChatGPT can help identify potential risk factors facing an organisation, identify gaps or inconsistencies in risk assessments, highlight areas of the RM process that require further attention or improvement, and provide recommendations for risk response strategies based on past performance. Much of the current discussion surrounding applications like ChatGPT is concerned with the question of how well it works now and in the future. We believe that this question needs to be approached with a clearly defined task in mind. To the best of the authors' knowledge, no research has been conducted to evaluate the performance of ChatGPT against any RM standards or professional guidelines. Therefore, this research aims to evaluate the performance of ChatGPT for the RM process based on ISO 31000 standard using expert evaluation.

2. Research Methodology

The research employed a mixed-methods approach for data collection, analysis and processing. For data collection, the authors conducted a focus group session with construction experts in Iraq who are specialised in construction RM. The focus group method was selected for qualitative data collection due to its ability to collect data from multiple participants simultaneously, providing diverse perspectives and insights for exploratory research (Nyumba et al. 2018). In addition, the interactive nature of focus groups can stimulate discussion and generate new ideas or insights in areas with limited existing knowledge. Focus groups usually require small groups of participants, of about 6 to 12 people per group (Harthi, 2015). In this research, one focus group session was conducted with nine experts working in the Iraqi construction industry in order to: (1) identify the key indicators for measuring the performance of ChatGPT in managing risks based on ISO 31000; and (2) determine the key assessment criteria for evaluating the identified indicators.

Following the focus group session, the participants were invited to conduct a user testing experiment in which they were asked to plan RM activities for a construction project, identify a set of 30 construction risks, estimate the probabilities and impacts of the identified risks, suggest response strategies for each risk factor, and establish a plan for risk communication, monitoring and controlling. Based on each user experience, the authors distributed a questionnaire survey to them to assess the identified ChatGPT performance indicators based on the Ghat GPT outputs using a five-point Likert scale ranging from 0.1 "very low" to 0.5 "very high". The questionnaire survey was chosen as the preferred method for data collection for several reasons. Firstly, it enables respondents to provide information in a structured and standardised manner, which facilitates data analysis and comparison. Secondly, it provides a degree of anonymity and confidentiality, which may encourage respondents to be more honest and open in their responses. The authors employed manual content analysis for both the outputs of focus group sessions and survey responses. In content analysis, key information is extracted from verbal, written, or video files, either quantitatively or qualitatively (Krippendorff, 2018). This method is highly effective in organising and analyzing information within documentary data and has been employed extensively in previous construction engineering and management research. In this research, the key factors were identified and sorted in a constructive way during the analysis process.

For data processing, the authors employed fuzzy set theory, a mathematical framework designed to handle uncertainty and imprecision (Al-Mhdawi, 2020). In the context of processing survey outputs, fuzzy set theory can be used to handle the imprecision and uncertainty that may arise in the data (Al-Mhdawi et al., 2023; Radhika and Parvathi, 2016). For example, in a Likert scale survey where participants rate their agreement with a statement on a scale of 1 to 5, there may be ambiguity in the interpretation of responses that fall in between the scale points (e.g., a response of 3.5). The authors developed a fuzzy-based assessment model to quantify the level of significance of ChatGPT performance indicators for the ISO 31000 standard. The developed model was constructed employing three primary processes: fuzzification, fuzzy inference, and defuzzification. The following sub-sections provide details for each process.

Fuzzification

Fuzzification is a fundamental process in fuzzy set theory that involves transforming crisp or deterministic data into fuzzy or uncertain data. In other words, fuzzification is the process of mapping precise numerical values or discrete states onto fuzzy values, which are represented by membership functions. The current study utilised the triangular membership function for this purpose. This method is commonly used to represent fuzzy sets due to its simplicity and effectiveness (Al-Mhdawi et al., 2022a; Nayak et al., 2020). The triangular membership function is characterised by three parameters: the left, centre, and right values, which determine the location of the function's peak and the width of the function. Triangular membership functions have proven to be particularly useful in capturing subjective and imprecise information, and they offer the advantage of allowing for easy definition of the input range and straightforward arithmetic calculations (Sadollah, 2018; Al-Mhdawi et al., 2023).

Fuzzy interference

Fuzzy inference is the process of formulating the mapping from a given input to an output using fuzzy logic. In this study, the Mamdani fuzzy inference system (MFIS) was used to assess the output variable. The MFIS is one of the most widely used fuzzy inference systems. It uses if-then rules that relate the input variables to the output variables (Kaur and Kaur, 2012). The MFIS also has an intuitive nature that makes it easy for experts to interpret and use, and it is particularly suitable for subjective inputs that are difficult to quantify (Al-Mhdawi et al., 2022b). In this research, a total of 125 if-then rules were employed. The rules were developed based on prior work (see e.g., Al-Mhdawi et al. 2023).

Defuzzification

Defuzzification is the process of converting the fuzzy output of a fuzzy inference system back into a crisp or numerical value that can be used as a decision or an action. In this study, the centroid of area method was employed for this purpose. The centroid of area method is a widely used method of defuzzification that reflects the viewpoint of the experts. It calculates the centre of mass of the fuzzy set by taking the weighted average of the fuzzy values or membership grades. The result is a crisp value that represents the centre of the fuzzy set and can be used for further analysis or decision-making (Nieto-Morote and Ruz-Vila, 2011).

3. Results and Discussion

Profiles of Focus Group Participants

The session's members were experts in managing construction projects risks within the Iraqi construction industry. Those experts were contractors, project managers, and academics. The profiles of the participants are presented in Table 1.

Identified Key Performance Indicators and their Quantification

In total, 12 indicators were identified to measure the performance of ChatGPT for managing construction risks based on ISO 31000 standard. The identified performance indicators and their descriptions are presented in column 1 and 2 of Table 2. Following the focus group session, user experience testing was utilised to evaluate the performance of ChatGPT for the ISO 31000 RM process. Two assessment criteria were employed to evaluate ChatGPT's performance indicators: Validity (*v*) and Scalability (*s*). Validity refers to the accuracy and reliability of the performance indicator in measuring what it is intended to measure accurately and reliably. Scalability refers to the ability of the performance indicator to be applied to different levels of the organisation, from individual departments to the entire organisation.

As mentioned in the research methodology, the ranking was conducted using a questionnaire survey and processed under a fuzzy environment. The architecture of the proposed performance assessment model consisted of one fuzzy controller. The inputs were "*v*" and "*s*" and the output variable was Fuzzy Performance Number (*fpn*). The fuzzy input-output relations were determined using 125 IF-THEN rules. The quantified performance indicators of ChatGPT and their corresponding rankings are displayed in columns 3 to 6 of Table 2. Figure 1 illustrates the architecture of the proposed ChatGPT performance assessment model.

Table 1. Profiles of participants in the focus group session

Number of Group Participants	Construction Role	Range of Experience	Education Level			
			Dip	**BSc**	**MSc**	**PhD**
5	Project managers	15-21	1	3	1	-
2	Construction contractors	18-24	-	2	-	-
2	Academics	13- 16	-	-	-	2

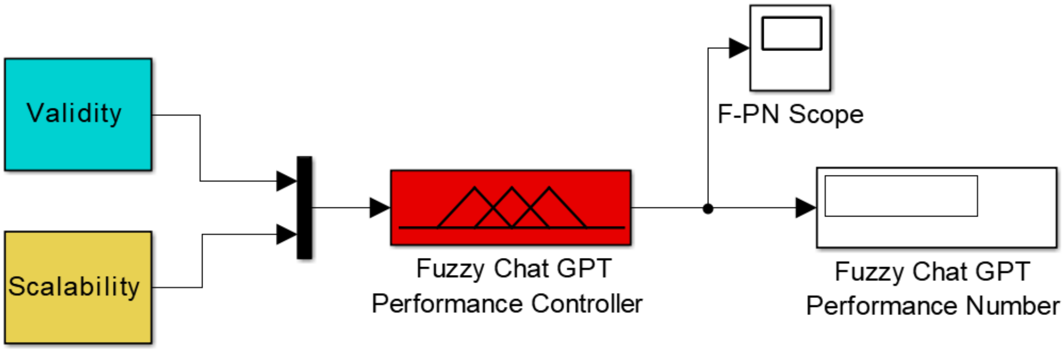

Figure 1. The proposed ChatGPT performance assessment model

Table 2. Identified ChatGPT performance indicators, their descriptions, and quantification.

ChatGPT Performance Indicators	Description	ChatGPT performance ranking			
		v	***s***	***f-pn***	***Rank***
P01. Accuracy of risk identification	It refers to the extent to which ChatGPT can accurately identify potential risks that may affect the construction project.	0328	0.393	0.401	7
P02. Ability to provide relevant risk mitigation strategies	It refers to the extent to which ChatGPT can suggest effective risk mitigation strategies that are relevant to the construction project.	0.388	0.370	0.447	1
P03. Consistency of responses	It refers to the extent to which ChatGPT provides consistent responses to similar questions or inputs.	0.369	0.293	0.374	10
P04. Clarity of communication	It refers to the extent to which ChatGPT can clearly communicate its responses to the user, including the language used and the level of detail provided.	0.411	0.361	0.431	2
P05. Ability to learn and adapt to new information	It refers to the extent to which ChatGPT can learn from new information and adjust its responses accordingly.	0.328	0.370	0.392	8
P06. Consistency in risk assessment and prioritisation	It refers to the extent to which ChatGPT consistently assesses and prioritizes risks in a way that is aligned with the construction project's objectives.	0.319	0.339	0.347	12
P07. Flexibility to customise risk management processes	It refers to the extent to which ChatGPT outputs can be customized to fit the specific needs and goals of the construction project.	0.291	0.393	0.382	9
P08. Compliance with industry standards and best practices	It refers to the extent to which ChatGPT's risk management processes are in line with industry standards and best practices for the construction industry.	0.300	0.357	0.360	11
P09. Continual improvement and updates	It refers to the extent to which ChatGPT is continually updated and improved to address new risks and emerging trends in the construction industry.	0.337	0.398	0.408	5
P10. Ability to handle multi-language input	It refers to the extent to which ChatGPT can handle input in different languages, which may be useful for international construction projects.	0.351	0.402	0.421	4

Table 2. *Continued*

ChatGPT Performance Indicators	Description	ChatGPT performance ranking			
		v	***s***	***f-pn***	***Rank***
P11. Compatibility with different devices and platforms	It refers to the extent to which ChatGPT is compatible with different devices and platforms, such as desktop computers, mobile devices, or cloud-based platforms.	0.360	0.416	0.428	3
P12. Ease of use	It refers to the extent to which ChatGPT is user-friendly and intuitive, enabling project team members to easily use the tool for risk management purposes.	0.346	0.439	0.406	6
		***Mean f-pn* = 0.3997**			

The quantified performance of ChatGPT was evaluated on a scale of 0.1 to 0.5, where 0.1 indicates very poor performance and 0.5 represents very high performance of ChatGPT. The mean value of Chat GPT's *f-pn* was found to be 0.3997, indicating a high level of performance in managing risks.

The evaluation revealed that the highest performance area of ChatGPT was found to be in its ability to provide relevant risk mitigation strategies. This is a critical component in RM, as it helps organisations to identify and implement effective measures to reduce or mitigate risks. ChatGPT's ability to provide relevant strategies was likely due to its ability to analyse large volumes of data and identify patterns and trends that human experts may not have been able to identify. However, the research also found that the consistency in risk assessment and prioritisation was the lowest performance area of ChatGPT. This means that ChatGPT may need further improvements in this area in order to consistently provide accurate risk assessments and prioritisation. This is a vital component of RM, as it enables organisations to determine the level of risk posed by different factors and prioritise their mitigation efforts accordingly.

4. Conclusions

This paper focuses on evaluating the performance of ChatGPT for the RM process based on the ISO 31000 standard by surveying construction and project risk academics in Iraq. In this research, the authors (1) identified the key indicators for measuring the performance of ChatGPT in managing construction risks based on ISO 31000, (2) determined the key assessment criteria for evaluating the identified indicators, and (3) developed a fuzzy-based assessment model to quantify the level of performance of ChatGPT by measuring the (*v*) and (*s*) of the identified performance indicators. The key conclusions are summarised as follows:

1. ChatGPT displayed a high level of performance in managing project risks under ISO 31000 standard.
2. The provision of relevant risk mitigation strategies by ChatGPT proved to be highly advantageous for risk management efforts. The platform's ability to suggest effective measures and solutions contributed significantly to improving risk outcomes.
3. Caution must be exercised when relying solely on ChatGPT for risk assessment and prioritisation. While the overall performance of the platform is high, it is still important to perform manual review and verification to ensure the accuracy and reliability of the risk assessments.
4. Further development efforts are necessary to enhance the accuracy and reliability of AI systems for risk management. Improving the precision and dependability of these systems will significantly enhance their effectiveness in identifying and addressing project risks.

To this end, this research serves as a foundation for future studies and developments in the field of AI-driven risk management, advancing our theoretical understanding of the application of AI models like ChatGPT in real-world risk scenarios. Further studies with larger sample sizes and more risk management tasks are recommended to validate the research findings.

Acknowledgement

The authors gratefully acknowledge the Iraqi construction experts for their valuable contributions to this study. Their participation, insights, and expertise greatly enhanced the research.

References

Al-Mhdawi, M. K. S. (2020). Proposed risk management decision support methodology for oil and gas construction projects. In *Proc., 10th Int. Conf. on Engineering, Project, and Production Management,* 407–420.Singapore: Springer

Al-Mhdawi, M. K. S., Brito, M., Abdul Nabi, M., El-adaway, I, and Onggo, B.S. (2022a). Capturing the impact of COVID-19 on construction projects in developing countries: A case study of Iraq. *Journal of Management in Engineering, 38* (1): 05021015. https://doi.org/10.1061/(ASCE)ME.1943-5479.0000991

Al-Mhdawi, M. K. S., O'Connor, A., Brito, M., Qazi, A., and Rashid, H. A. (2022b). Modeling the effects of construction risks on the performance of oil and gas projects in developing countries: Project managers' perspective. In *Proc*., *Civil Engineering Research in Ireland Conf. (CERI2022), Dublin, Ireland, 25-26 August 2022, Niall Holmes, Caitriona De Pbaor & Roger P. West, 2022*, 486–491. Dublin, Ireland: Civil Engineering Research in Ireland and Irish Transport Research Network.

Al-Mhdawi, M.K.S., Brito, M., Onggo, B.S., Qazi, A. and O'Connor, A. (2023). COVID-19 emerging risk assessment for the construction industry of developing countries: evidence from Iraq. *International Journal of Construction Management*, pp.1-14.

Harthi, A. (2015). Risk management in fast-track projects: a study of UAE construction projects [Doctoral dissertation]. University of Wolverhampton.

Kaur, A. and Kaur, A. (2012). Comparison of Mamdani-type and Sugeno-type fuzzy inference systems for air conditioning system. *International Journal of Soft Computing and Engineering, 2*(2), pp.323-325

Krippendorff, K. (2018). *Content analysis: An introduction to its methodology*. Sage publications.

Lalonde, C. and Boiral, O. (2012). Managing risks through ISO 31000: A critical analysis. *Risk management*, *14*, pp.272-300.

Nayak, S., Pattanayak, S., Choudhury, B.B. and Kumar, N. (2020). Selection of industrial robot using fuzzy logic approach. In *Computational Intelligence in Data Mining* (pp. 221- 232). Springer, Singapore

Nieto-Morote, A. and Ruz-Vila, F. (2011). A fuzzy approach to construction project risk assessment. *International Journal of Project Management, 29*(2), pp.220-231

Nikolic, S., Daniel, S., Haque, R., Belkina, M., Hassan, G.M., Grundy, S., Lyden, S., Neal, P. and Sandison, C. (2023). ChatGPT versus engineering education assessment: a multidisciplinary and multi-institutional benchmarking and analysis of this generative artificial intelligence tool to investigate assessment integrity. *European Journal of Engineering Education*, pp.1-56.

Nyumba, T., Wilson, K., Derrick, C.J. and Mukherjee, N. (2018). The use of focus group discussion methodology: Insights from two decades of application in conservation. *Methods in Ecology and evolution, 9*(1), pp.20-32.

Radhika, C. and Parvathi, R. (2016). Intuitionistic fuzzification functions. *Global Journal of Pure and Applied Mathematics, 12*(2), pp.1211-1227

White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J. and Schmidt, D.C. (2023). A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382*.