Can ChatGPT exceed humans in construction project risk management?

Roope Nyqvist
School of Engineering, Aalto University, Espoo, Finland, and
Antti Peltokorpi and Olli Seppänen
Department of Civil Engineering, Insinööritieteiden korkeakoulu, Aalto-yliopisto, Espoo, Finland

Received 14 August 2023
Revised 20 December 2023
18 February 2024
Accepted 22 February 2024

Abstract

Purpose – The objective of this research is to investigate the capabilities of the ChatGPT GPT-4 model, a form of artificial intelligence (AI), in comparison to human experts in the context of construction project risk management.

Design/methodology/approach – Employing a mixed-methods approach, the study draws a qualitative and quantitative comparison between 16 human risk management experts from Finnish construction companies and the ChatGPT AI model utilizing anonymous peer reviews. It focuses primarily on the areas of risk identification, analysis, and control.

Findings – ChatGPT has demonstrated a superior ability to generate comprehensive risk management plans, with its quantitative scores significantly surpassing the human average. Nonetheless, the AI model's strategies are found to lack practicality and specificity, areas where human expertise excels.

Originality/value – This study marks a significant advancement in construction project risk management research by conducting a pioneering blind-review study that assesses the capabilities of the advanced AI model, GPT-4, against those of human experts. Emphasizing the evolution from earlier GPT models, this research not only underscores the innovative application of ChatGPT-4 but also the critical role of anonymized peer evaluations in enhancing the objectivity of findings. It illuminates the synergistic potential of AI and human expertise, advocating for a collaborative model where AI serves as an augmentative tool, thereby optimizing human performance in identifying and managing risks.

Keywords Artificial intelligence (AI), Large language models (LLM), ChatGPT, GPT-4, Risk management, Construction management, Project management, Risk analysis

Paper type Research paper

1. Introduction

The radical emergence of Artificial Intelligence (AI) has initiated a paradigm shift across multiple industries, recalibrating conventional modes of operation and harnessing innovative potential (Chui *et al.*, 2023; Eloundou *et al.*, 2023). One emerging beneficiary of this technological evolution is the construction industry, which has historically been associated with volatility, unpredictability, and inherent risk (Smith *et al.*, 2006; Project Management Institute, 2016).

We thank the companies and experts involved in this research for their valuable feedback and their cooperation with the research. Furthermore, AI tools were used in a responsible way, as described prior, and to support the writing of the article.

Engineering, Construction and Architectural Management
Vol. 31 No. 13, 2024
pp. 223-243
Emerald Publishing Limited
0969-9988
DOI 10.1108/ECAM-08-2023-0819

In light of these challenges, the concept of construction project risk management (CPRM) has long played a key role in ensuring project success. Modern CPRM methods require intricate planning, systematic assessment, and comprehensive control measures (Project Management Institute, 2016; International Organization for Standardization, 2018). However, despite advances, traditional CPRM practices often fall short in their efficiency and accuracy, a void that AI promises to fill.

Recent research indicates that AI's application in construction project management, particularly in CPRM, could significantly enhance efficiency and accuracy, pointing to a new era of advancements (Abioye *et al.*, 2021). However, comprehensive investigations into AI's role in CPRM are still rising (Zhao, 2022).

Previous studies, often using less advanced models like GPT-3.5, show varied results and potential biases due to non-anonymous author evaluations or focus groups (Aladağ, 2023; Barcaui and Monat, 2023; Hofert, 2023). This highlights the necessity for more objective methods, such as blind reviews, to validate AI capabilities in this domain (Shoham and Pitman, 2020; Monaghan *et al.*, 2021). The majority of existing literature pre-dates significant developments like ChatGPT's June 2022 and GPT-4's March 2023 releases (Abioye *et al.*, 2021).

This study aims to fill this research gap, exploring the ability of OpenAI's GPT-4-based model, ChatGPT, to outperform human experts in CPRM. The GPT-4 model's ability to understand and generate human-like text (OpenAI, 2023) provides a unique lens to examine the potential for AI applications in risk management. Additionally, this study utilizes a CPRM test, with anonymous blind peer-reviews to determine AI and human capabilities, enhancing the objective comprehension of AI capabilities in CPRM.

The objective of this study was to explore the capabilities of AI, specifically ChatGPT (GPT-4), in performing CPRM with a focus on risk identification, risk analysis, and risk control compared to human professionals (International Organization for Standardization, 2018). To achieve this objective, we designed an empirical skills test that was used to gather responses both from ChatGPT (GPT-4) and human experts through a series of CPRM-related questions. These tasks revolved around a simulated case project, providing a unique dataset that could be used for a comparison of generative AI and human performance in CPRM through an anonymous peer review process.

Interestingly, the results of our research revealed indications of the superiority of ChatGPT over human experts in the test comparison. This finding undermines the narrative of human dominance in complex decision-making tasks and reveals the potential of generative AI to support traditional CPRM processes alongside human experts.

The implications of this study are profound and far-reaching. Our research suggests that AI solutions, such as ChatGPT (GPT-4), can serve as powerful tools to augment and refine existing risk management protocols. This use indicates the potential to bring greater precision, efficiency, and effectiveness to the process, enabling more informed decision-making and risk control when its limitations are properly addressed.

This paper serves as both a report of the research findings and a call to the building industry to further investigate the practical uses of AI in construction management. The research uncovers AI's abilities in the CPRM compared to human professionals through anonymous peer review and challenges the status quo by offering insights into a new AI-powered risk management approach to build a safer, more predictable, and less risky future.

2. Literature background

The past few decades have witnessed the remarkable rise of artificial intelligence (AI) from the realm of science fiction to the heart of numerous professional fields (Chui *et al.*, 2023). Its ability to learn, adapt, and emulate human cognitive functions has evolved exponentially,

driven by advances in machine learning, neural networks, and natural language processing (Eloundou *et al.*, 2023; OpenAI, 2023). The potential for AI to revolutionize traditional industries is vast, and its impact on improving efficiency, precision, and productivity has been lauded by businesses (Chui *et al.*, 2023) and academic circles (Abioye *et al.*, 2021; Bolpagni *et al.*, 2021).

At the same time, the construction industry, which has often been criticized for its resistance to innovation, faces persistent challenges (Liu *et al.*, 2018; Akinosho *et al.*, 2020; Bolpagni *et al.*, 2021). In particular, risk management in construction projects remains fraught with uncertainty due to complexities related to various aspects such as logistics, planning, safety, budget, and schedule constraints (Project Management Institute, 2016). Traditional risk management methods, such as checklists, judgmental heuristics, and sensitivity analysis, have been unable to fully mitigate these risks due to their inherent limitations (Chapman and Ward, 2011; Maldonato and Dell'Orco, 2011).

Moreover, challenges in managing uncertainty and risk significantly have hindered the effectiveness of project management. This issue is reflected in the high rate of project failures; surveys reveal that 37% of projects have significantly missed their budget and or schedule performance targets due to lack of effective risk management (Armstrong *et al.*, 2023). Furthermore, particularly large-scale projects often run up to 80% over budget and exceed their scheduled timelines by 20% (Agarwal *et al.*, 2016).

This trend is not confined to a specific sector; it spans various industries, with 12% of projects classified as failures and 65% not meeting their scheduled completion times (Project Management Institute, 2021). The statistics underscore the difficulty in predicting and controlling project variables, highlighting the need for more effective project risk management.

To address the aforementioned problems, various CPRM methods have been developed, such as deep learning (Akinosho *et al.*, 2020), Bayesian network-based methods (Arabi *et al.*, 2022), and uncertainty network modeling (Nyqvist *et al.*, 2024). However, the existing methods have not yet been able to significantly transform the construction industry performance, and performance problems persist (Agarwal *et al.*, 2016; Project Management Institute, 2021).

However, novel technologies around generative AI have emerged that indicate potential to address the challenges. Research has highlighted the capabilities of novel AI technologies (e.g. ChatGPT, Gemini, and Llama) in their ability to analyze extensive datasets, identify complex patterns, and generate insights (OpenAI, 2023; Touvron *et al.*, 2023, Google, 2023), indicating potential also in application within CPRM.

Therefore, the intersection of AI and CPRM provides a fertile area for research and innovation. A handful of studies have examined the application of AI to specific tasks in the construction domain, such as planning, scheduling, and cost estimation (Abioye *et al.*, 2021; Zhao, 2022; Aladağ, 2023; Rane, 2023).

Generally, the early work using AI in construction risk management has been promising, highlighting the technology's ability to identify, analyze, and mitigate risks more effectively than traditional methods (Abioye *et al.*, 2021; Kamari and Ham, 2022).

In a recent study, Aladağ (2023) used non-anonymous expert focus groups to determine the accuracy of ChatGPT (GPT-3.5 demo version) in CPRM. The findings indicated a moderate level of performance in managing risks. Furthermore, outside of construction projects, ChatGPT proficiency in risk management has been investigated with generally positive, but mixed results through author evaluation (Barcaui and Monat, 2023; Hofert, 2023).

These studies suggest that AI can add a layer of sophistication and accuracy to risk management. However, research on large language models and generative AI, such as ChatGPT, has been scarce in the CPRM context (Abioye *et al.*, 2021; Zhao, 2022; Aladağ, 2023)

and a full revelation of the potential and limitations of AI for CPRM has been largely uncovered (Klepo *et al.*, 2023) as the majority of publication predate the publication of ChatGPTs GPT-3 2022 and GPT-4 2023, indicating a that the general novelty of the research topic persists.

The existing studies are not comprehensively representative of the GPT-4 model capabilities in CPRM (OpenAI, 2023), with authors such as Aladağ (2023) and Hofert (2023) utilizing the less powerful GPT-3.5 model. Additionally, the used research methods pose potentially biased results, due to reliance on author evaluations (Barcaui and Monat, 2023; Hofert, 2023), or non-anonymous focus groups evaluations (Aladağ, 2023) indicating a need for additional anonymous reviews to confirm the findings (Shoham and Pitman, 2020; Monaghan *et al.*, 2021).

Therefore, despite the growing progress of AI applications in the construction industry (Abioye *et al.*, 2021; Zhao, 2022, Ghimire *et al.*, 2023), the full extent of its transformative potential has remained shrouded in uncertainty, posing a persisting knowledge gap on the measurement of AI capabilities. The general understanding of how generative AI solutions, such as ChatGPT, could be used to revolutionize CPRM is still insufficient. A deeper dive into this topic could provide a more nuanced understanding of the potential practical and managerial applications and limitations of AI.

In conclusion, the confluence of AI and CPRM represents an exciting frontier for research and innovation. Through rigorous exploration and testing, the industry can uncover the transformative potential of AI and capitalize on the opportunities it presents. In doing so, we can shift the paradigm of CPRM and create a more resilient, efficient, and forward-looking industry. In the following sections, this study aims to contribute to this expanding discourse and shed light on the potential of generative AI-powered solutions in CPRM.

3. Research design and methods

This study used a mixed-methods approach pitting the AI solution, ChatGPT GPT-4, against human experts in the field of CPRM through a test, where participants answer a series of questions related to the risk management of a simulated case project. After answering, all of the responses were anonymized and peer-reviewed by other research participants to provide data that was then analyzed by the lead author, both qualitatively and quantitatively, to provide answers to the research question "Can ChatGPT exceed humans in construction project risk management?"

Blind peer review was chosen to provide an objective approach because of its robust reliability in areas such as scientific publishing (Shoham and Pitman, 2020) and medical testing (Monaghan *et al.*, 2021).

This research process, as illustrated below in Figure 1, was developed to ensure the validity and reliability of the findings, the research paid careful attention to the design and validation of the test questions. The goal was to create a robust measure of risk management capability close to a real-world situation, where bias towards or against AI or human responses could be mitigated through anonymous peer review.

In Phase 1, human participants were selected from different Finnish construction companies, reflecting a diversity of experiences and perspectives while being able to test the advantages of localized knowledge in the following test to indicate AI capabilities in a simulated real-world case.

As the AI participant, ChatGPT GPT-4 was chosen, as it functions as a generative AI, capable of generating text based on the user's prompts, and due to its leading position in the field of generative AI solutions during the research conducted in the spring of 2023.

The focus of the research was a case project carefully designed to emulate the real-life situation and complexities encountered in CPRM to showcase human and AI capabilities in a

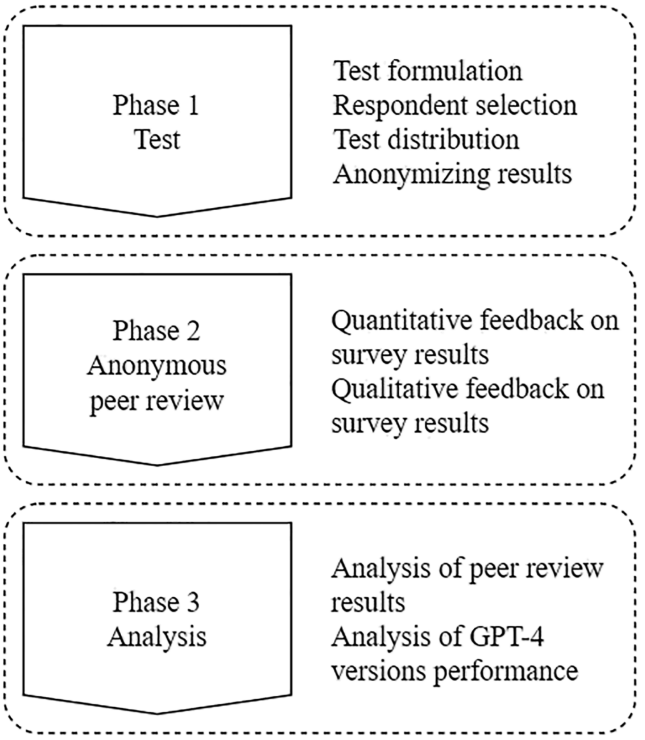

Figure 1.
Research process

valid, realistic setting. The case project was tailored to include region-specific information and to replicate scenarios where local experts (e.g. human participants) might have an advantage. The full case project description goes as follows: You are the main contractor in a renovation project of a medium-sized health center in Helsinki. The site is located in a densely populated area in Töölö. The building was originally built in the 1930s but has been renovated several times during its life cycle. The area to be renovated covers an area of approximately 4,500 m2. The renovation project is being commissioned by a financially sound company with which you have a fixed-price contract for the renovation work. The contract stipulates that the entire renovation work is to be carried out in 2024. A special feature of the project is that part of the premises will remain in use as a health center during the project.

The case project description was created to simulate a real construction project with specific details such as construction type, location, size, age of the building, contract type, client details (i.e. financial status), building use, and special constraints associated with the project. The case characteristics were chosen to encourage responses that capture the real-world complexities and intricacies of CPRM, as opposed to generic responses.

Next, in Phase 2 the test was conducted around the case project, with both human experts and ChatGPT answering a series of questions about the case project, through an online questionnaire. These included the following series of questions: First, respondent information, including name, position and years of experience. Second, identify and list what are the potential risks to the project? Third, which of these risks are the most critical? Could you analyze them? Third, tell us what are you doing to control these critical risks?

The questions were designed to simulate three key risk management steps, risk identification, risk analysis, and risk control. The baseline data provided to answer these questions was identical for both the human experts and ChatGPT.

The study generated a total of 16 responses from the human experts, initially chosen to respond based on their relevant experience towards CPRM (see Table 1 for respondent information including respondent titles, and experience), and one AI response by ChatGPT using the GPT-4 model. The test introduction, case descriptions, and all of the responses were generated in Finnish to fit the human respondent's native language.

Upon receipt of the responses, all were anonymized to ensure confidentiality and then redistributed among the participants. Each participant received a package containing two randomly selected human responses, excluding their own, and one response generated by ChatGPT. The participants were not given information on whether responses were given by humans of ChatGPT. All participants had given informed consent for their responses to be used and were assured of their anonymity throughout the process.

A total of 19 voluntary human expert reviewers and ChatGPT GPT-4 participated in the evaluation. Each of the reviewers was asked to provide quantitative scores on a scale of 1–10 on individual responses, an overall score, and qualitative feedback on each of the responses they reviewed. This provided a consistent scoring framework for all reviewers and resulted in quantitative and qualitative data for further analysis.

Finally, in Phase 3 the data analysis included both quantitative and qualitative methods. For the quantitative analysis, the research calculated the average and standard deviation of the scores. The qualitative feedback was analyzed by grouping the feedback into positive and negative, and by identifying recurring patterns and insights to create a thematic synthesis of the data.

Throughout the study, participants were unaware of ChatGPT participation, or details of the other human participants to protect the study from potential bias. One of the reviews submitted only qualitative feedback, and five reviewers did not provide any qualitative feedback.

Strict measures were taken to manage and store the collected data securely. All data was anonymized and access was restricted to the research team to ensure participant confidentiality. Respondents were not told which responses were given by AI and which by humans to reduce potential bias towards answering the questions.

Table 1. Respondents information

Response	Title	Experience in years
Human response 1	Project manager	10
Human response 2	Site manager	20
Human response 3	Director of development	25
Human response 4	Site officer	6
Human response 5	Production engineer	5
Human response 6	Head of project management office	38
Human response 7	Risk manager	4
Human response 8	Safety manager	36
Human response 9	Production coach	7
Human response 10	Project manager	22
Human response 11	Quality manager	8
Human response 12	Head of technical office	15
Human response 13	Project manager	39
Human response 14	Project manager	27
Human response 15	Risk manager	8
Human response 16	Project manager	20
ChatGPT response	Artificial intelligence	

Finally, the authors re-tested ChatGPT's GPT-4 responses on the original test and made a qualitative comparison between the AI's responses from April 2023 and December 2023 to determine the variation in response quality and whether it could affect the overall assessment of ChatGPT's capabilities.

4. Analysis and results

The structure of the analysis and results presentation is divided into two parts: a quantitative analysis and a qualitative analysis. The quantitative section compares the performance of ChatGPT and human experts through a scoring system. The qualitative portion evaluates the feedback provided by human experts on risk management strategies derived by both humans and ChatGPT. It concludes with a summary of the strengths and weaknesses of generative AI-driven risk management, underlining the potential and challenges of AI in this field through an anonymized peer-reviewed test.

4.1 Quantitative analysis and results

ChatGPT achieved significantly better quantitative evaluations than humans both when reviewed by human reviewers and ChatGPT itself. Human experts average 5.7, while ChatGPT averages a score of 8.6 (see Table 2) on the exercise based on human review.

Based on human reviews	
Sample size (n)	18
Summary	
Human average	5.7 ± 1.9
ChatGPT average	8.6 ± 1.2
Human capability	
Risk identification	6.1
Risk analysis	5.3
Risk control	5.8
ChatGPT capability	
Risk identification	8.2
Risk analysis	8.4
Risk control	8.6
Based on ChatGPT review	
Sample size (n)	1
Summary	
Human average	7.6 ± 1.0
ChatGPT average	9.0 ± 0.5
Human capability	
Risk identification	7.9
Risk analysis	7.4
Risk control	7.4
ChatGPT capability	
Risk identification	10.0
Risk analysis	9.0
Risk control	8.0

Table 2. Summary of human and ChatGPT capabilities

None of the human respondents' total average score exceeded the ChatGPT average score. Overall ChatGPT had a smaller standard deviation than humans. The exhaustive quantitative scoring can be seen in Appendix.

Also, ChatGPT was used to review all of the responses, including its own. It scored human respondents more positively than other humans. Furthermore, it scored its answer higher on the risk identification and risk analysis parts, but humans valued the AI answers to risk control higher than ChatGPT.

From the results, it can be seen that ChatGPT can exceed the average human experts participating in this anonymized test. Furthermore, none of the human respondents disclosed using AI as a tool to help them answer the exercise.

4.2 Qualitative analysis and results

The analysis conducted was to compare and contrast the written feedback provided by human experts and ChatGPT on each respondent's answers to the test. The research used a set of evaluative comments provided by human experts as data to assess the capabilities of both human and ChatGPT-driven risk management. The evaluative comments were provided in response to both the human and ChatGPT responses to the same test to achieve appropriate comparability.

During the comparison, the research found a variety of responses among the evaluators to both the human and ChatGPT responses. The evaluators' translated feedback was first divided into positive and negative feedback. Second, the main author conducted a further thematic classification of the data, as seen in Table 3, where Rn is the reviewer and Vn is the response number. For additional clarity, ChatGPT GPT-4 responses are marked with the abbreviation AI.

The first commonality the research noticed was the human evaluators' emphasis on the comprehensiveness of risk management plans. Evaluators emphasized the importance of considering as many critical risks as possible and providing plans for managing them. Some evaluators praised ChatGPT's ability to provide comprehensive and detailed risk assessments, suggesting that generative AI systems have the potential to excel in this area.

However, an equally common concern was the specificity and practicality of risk management. Human evaluators criticized both human and AI-generated responses for failing to tailor their risk management strategies to the specific circumstances of the construction project. While the AI responses were praised for their breadth, they were criticized for being too general and lacking actionable strategies indicating a need for human experts' tacit knowledge.

Another point of contention was the feasibility of implementing the strategies proposed by AI. Some evaluators noted that the proposed measures seemed difficult to implement in practice. This highlights a potential limitation of AI systems in risk management, as their solutions may not always take into account the practical constraints and limitations inherent in construction projects.

Interestingly, some evaluators expressed concern about the too-detailed approach to risk management presented by ChatGPT, suggesting that the abundance of information could lead to a loss of focus or overwhelm project stakeholders.

In contrast to other respondents, one outlier reviewer offered a significantly lower rating (see human reviewer 19 in Appendix) and articulated dissatisfaction with the general level of responses made by ChatGPT (without knowledge of AI as a respondent). The reviewer highlighted a lack of recognition of the project's unique characteristics, an omission of critical risks such as the usability of the workspace and the safety of off-site personnel, and an inundation of information on general construction risks without addressing the project's unique risks.

Table 3. Thematic analysis of the qualitative data from human expert feedback on the responses

Theme	Positive feedback	Negative feedback
Comprehensiveness	R5 (V6): Risks identified from a variety of perspectives and broken down into different areas. All also identified a number of measures. At this level and with this information, a comprehensive assessment has been made	R1 (V4): The answers and analysis of risks are very limited and leave most of the potential risk factors unmentioned/unanalyzed
	R7 (V13): Respondent lists the three most significant risks. I agree with these. These have not been analyzed in detail in this section, but on the other hand respondent 13 has already opened up these issues in Section 2. For example 10 sentences per risk	R13 (V15): Key issues identified, but a limited response
	R17 (V1): The user is well highlighted, with small things (such as information, e.g. weekly newsletters) the project can be streamlined considerably. An angry user can put the whole project in crisis. Realities taken into account, updates to plans will always come. In scheduling, consideration of the collection of planning baseline data and changes is essential	R19 (V1): Identifies the features of the project but not the risks associated with that feature
	R1 (AI): Responses are comprehensive and take into account the majority of critical risks and their management in the type of site presented	R7 (AI): A broad sample of risks is a good thing in itself. Some of the risks were quite generic, it seems that a generic list of project risks was used
	R3 (AI): The answer is comprehensive and I think quite good	R19 (AI): Responses at a general level. No specific characteristics of the project are identified. Comprehensive list otherwise
	R5 (AI): Risks are identified from a variety of perspectives and broken down into different areas. All also identified a number of measures. At this level and with this information, a comprehensive assessment has been made	R19 (AI): Lack of identification of risks
	R7 (AI): The main risks are well-listed and analyzed. Clearly the best answer to point 3	R19 (AI): Did not open up the risks of the project
	R13 (AI): The response lists a wide range of risk types and risks	
	R17 (AI): Risks considered comprehensively from different perspectives, but at a general level	

(*continued*)

Table 3.

Theme	Positive feedback	Negative feedback
Specificity and practicality	R7 (V13): A broad sample of risks and well-opened them. Clearly the risks have been thought through specifically for this project. Clearly the best risk coverage of the responses presented	R3 (V1): In addition, neighbors, site organization and safety should be taken into account when working in a densely populated area in Töölö
	R8 (V11): Relevant risks identified, analyses in previous section but could be more comprehensive	R6 (V14): Answer confusingly in other paragraphs
	R8 (V11): This was a very explicit assessment of site-specific risks based on baseline data	R7 (V4): I would have liked to have seen a more detailed analysis of the risks, now I am left with general remarks
	R7 (AI): Very well opened up to more specific measures per risk. Clearly the best answer to point 4	R13 (V8): The response was limited and project specificity was not taken into account
	R9 (AI): In response, risk identification and analysis is broad but sufficiently focused on the project. The answers to question 2 are a bit more circular than those from V13, hence the lower score. Overall, the best answer, where the analysis is of high quality and clarifies the risks identified, followed by management measures that provide good steps based on the analysis	R14 (V1): The sentences are not complete, and for the most part leave it unclear how the potential risks presented will affect the achievement of the project's objectives
	R13 (AI): Critical risks are well selected and the main points are disclosed	R3 (AI): A few points that are left floating in the air without concrete action, e.g. "Use a procurement strategy that ensures competitive prices and quality materials"
	R17 (AI): Risks considered comprehensively from different perspectives, but at a general level (perhaps more specificity on some points?) The answers to questions 2 and 3 are general, but very concrete suggestions are given in answer 4	R8 (AI): Site specificity had been taken into account, but for the most part this was a fairly general list of risks that apply to any project
		R10 (AI): Clearly knows a lot about risk management, but the concrete measures risk getting lost in the mass of text/prioritization falls by the wayside

(*continued*)

Theme	Positive feedback	Negative feedback
Feasibility of implementation	R17 (V8): Concise answer, not much to evaluate. Occupational safety risks good points! Essential points mentioned, such as output data risks	R3 (V10): The answer is a bit off track. There is no mention in the brief of any zoning change or of the building being protected
	R19 (V1): Solution-oriented and project-specific	R8 (V13): A rather superficial answer
	R7 (AI): Very well opened up to more specific measures per risk	R3 (AI): There are certain points in the response which are extremely difficult to implement in practice
	R9 (AI): Overall, the best answer, where the analysis is of high quality and clarifies the risks identified, followed by management measures that provide good steps based on the analysis	R6 (AI): I would have liked to have seen some mention of proper resourcing of the site and possible double-shift working to manage the schedule risk
	R13 (AI): The issues are presented well and clearly. The process for dealing with surprising isolated anomalies and incidents is not specified, but at a higher level the answers are good	R11 (AI): Respondent is not in a position to lead these projects
	R19 (AI): A good attempt to prevent the risks identified. Partly describing the basic tasks, could describe the specificities of the project and the deviating measures	R15 (AI): The timetable given is so tight that it cannot be flexible – focus on advance planning and preparation
Abundance of information	R10 (V7): This is probably my favorite when you keep the answers suitably short (although not as comprehensive as respondent 17 [AI])	R7 (V4): In the big picture, significant risks identified, but remained at a rather general level. I would have liked to see a broader list of risks and a slightly more defined set of risks
	R10 (V14): It must be a genuine doer since he doesn't talk nonsense and the answers show the background of the construction	R19 (V9): Disruption management well identified, safety at work not. Aiming to do the job well is not a measure of risk management
	R15 (V16): Right things	R10 (AI): Clearly knows a lot about risk management, but the concrete measures risk getting lost in the mass of text/prioritization falls by the wayside
	R8 (AI): Essential risks identified from a long list and concisely but well analyzed	R15 (AI): I would have emphasized the specificities of renovation in terms of schedule and user cooperation, which also have an impact on the schedule, rather than quality risks
	R9 (AI): Overall, the best answer, where the analysis is of high quality and clarifies the risks identified, followed by management measures that provide good steps based on the analysis	
	R14 (AI): The respondent has identified in detail and in a comprehensive manner the potential challenges related to the project	

Table 3.

In summary, the qualitative analysis reveals a complex interplay of strengths and weaknesses in generative AI-driven risk management as embodied by ChatGPT. While ChatGPT demonstrates an impressive ability to comprehensively analyze and present risks within the test set, it also shows a potential gap in providing practical, context-specific, easily implementable strategies and seemingly lacks the implicit knowledge some human respondents could showcase. Nevertheless, the potential of AI in risk management seen through the anonymous peer-review results is significant.

4.3 Synthesis of results

The research attempts to answer the question, "Can ChatGPT exceed humans in construction project risk management?" To answer this question, the research used a mixed-methods approach that included both qualitative and quantitative analyses.

According to the quantitative results, the analysis of the sixteen human responses reviewed by nineteen different expert raters showed a significant difference between human and generative AI (ChatGPT GPT-4) capabilities. The average score from anonymous peer review of the human responses, 5.7, was significantly lower than that of ChatGPT, which was 8.6. This indicates a clear superiority of ChatGPT's capabilities over the average human expert participating in this study. The distribution of reviews was randomized and anonymous, and reviewers volunteered to rate a random selection of responses without knowing each other to ensure impartiality.

It is also noteworthy that ChatGPT was used to evaluate all human answers as well as its own. ChatGPT's ratings of human responses were more favorable than those of human reviewers, although it still rated its responses as superior on average. Interestingly, human reviewers rated ChatGPT's risk control capabilities higher than ChatGPT's self-assessment.

The qualitative analysis, on the other hand, showed mixed ratings for both human and AI responses. Although ChatGPT's risk management plans were praised for their comprehensiveness, the evaluators criticized them for their lack of practicality and specificity to the given construction project. On the other hand, while the human responses were criticized for their lack of comprehensiveness, they offered more specific and potentially more implementable solutions.

The combination of these quantitative and qualitative findings provides a holistic perspective on the capabilities of both humans and ChatGPT in managing construction project risks based on a test that simulates a real CPRM scenario. Based on the findings it is clear that ChatGPT excels at providing a broad view of potential risks due to its data processing capabilities, while humans bring more practical, tacit knowledge and context specificity to the table than AI could showcase in the test.

It should be noted, however, that none of the human respondents reported using AI solutions to support their responses. Had AI been used as an assistive tool for the human respondents, it might have improved the human average score, potentially allowing it to surpass the performance of ChatGPT. Therefore, the issue is not one of AI replacing humans, but rather one of using AI capabilities to enhance human performance in managing construction project risks.

In conclusion, while ChatGPT demonstrates superior capabilities in providing comprehensive risk management plans according to the quantitative results, the practicality and specificity of these plans need further improvement. Therefore, AI models such as ChatGPT may not necessarily surpass human capabilities in managing construction project risks in their current state, but they offer promising potential to enhance human performance when used as complementary tools.

4.4 Analysis of GPT-4 version performance

In addition to the capability measurements introduced prior through anonymous peer review, the authors conducted a qualitative analysis between ChatGPTs GPT-4 responses gathered on the CPRM test between April 2023 and December 2023.

The version from April 2023, illustrates GPT-4's initial approach to CPRM. It presents a broad spectrum of risks, adopting a generalist viewpoint that encompasses a wide range of potential issues, though without significant depth in specific risk strategies.

The second version from December 2023, shows a slight evolution in the model's approach. While covering similar risks, this version exhibits variances in articulation and content. The differences are subtle, focusing on variations in risk detailing rather than a complete overhaul of strategy.

Comparatively, both versions demonstrate GPT-4's ability to consistently handle construction project risks. The main distinction lies in the refinement of details and the presentation of risk management strategies, indicating a gradual, rather than a drastic, enhancement in AI capabilities for specialized tasks.

5. Discussion

The primary objective of this research was to investigate and understand the capabilities of both humans and ChatGPT (GPT-4) in the context of CPRM, specifically exploring the question "Can ChatGPT exceed humans in CPRM?" as the existing research literature has been insufficiently covering the topic (Akinosho *et al.*, 2020; Abioye *et al.*, 2021; Aladağ, 2023).

Prior research, conducted before 2022, inherently lacks insight into ChatGPT, given its more recent publication. Furthermore, contemporary studies often omit the implementation of blind peer review (e.g. Aladağ, 2023; Barcaui and Monat, 2023; Hofert, 2023) and typically focus on earlier, less advanced versions of ChatGPT, such as GPT-3.5 (e.g. Aladağ, 2023; Hofert, 2023). This scenario represented a research gap, particularly in understanding and defining the capabilities of ChatGPTs latest iteration, GPT-4, in CPRM.

To address the research gap and research question, the research conducted a CPRM test pitting human experts against ChatGPT (GPT-4). Answers to a series of questions related to the case project were collected from both ChatGPT and humans. Participants anonymously peer-reviewed the responses. Both quantitative scoring and qualitative feedback were analyzed.

The key findings of this study revealed that, under the conditions of the test applied, ChatGPT could surpass the performance of human experts. Quantitative results provided significant evidence to suggest the potential of generative AI in this domain.

ChatGPT demonstrated its ability to provide comprehensive and detailed risk assessments, exceeding the average performance of human experts. Comparingly, the results differ from the findings by Aladağ (2023), where experts evaluated ChatGPT (GPT-3.5 demo version) answers as moderate, but provided similar capabilities on word-based answers as research by Barcaui and Monat (2023) and Hofert (2023).

However, several factors could have influenced these results. First, this divergence can arguably be attributed, in part, to the influence of different ChatGPT versions on the performance of AI solutions, as discussed by Eloundou *et al.* (2023). Second, the application of an anonymous peer review method in this study reduces bias, in comparison to prior publications which have relied on small focus groups (Aladağ, 2023), or author assessment (Barcaui and Monat, 2023; Hofert, 2023).

As a result, the quantitative and qualitative results provide a new understanding of the capabilities of AI and humans on CPRM. The methods in this research arguably enabled a balanced evaluation of AI capabilities by paralleling comparisons in a similar test scenario, without participants having prior knowledge on the subject of assessment.

Furthermore, AI models are currently limited in their ability to access and interpret implicit data, such as personal experience with nuanced and complex underlying assumptions and hidden correlations, or unarticulated expertise and judgment, a dimension where human experts hold an advantage. This has been identified also in various literature and indicates advantages for human professionals in CPRM.

Compared to publications covering the capabilities of AI in other fields, this research found similar advantages. In previous publications, GPT-4 has performed well against humans in tests such as the Uniform Bar Exam, SAT math exams, and medical exams both in the United States and Spain (Eloundou *et al.*, 2023; Ray, 2023). Additionally, Barcaui and Monat (2023) found that AI-generated plans can introduce novel insights, especially in areas like risk management. Similarly, ChatGPT GPT-4 performed well in this research indicating that these generative AI solutions have widespread implications across different fields.

Additionally, the research provided insight into version differences between GPT-4 April 2023 and December 2023 versions. While the findings indicate that generative AI produces varying responses to the CPRM capability test, the variation is smaller than that between this study and contemporary literature utilizing GPT-3.5.

The findings were mostly consistent with the literature covered above. However, it could be concluded that the findings presented a more robust method of measurement, and improved the existing definitions of the capabilities of the ChatGPT GPT-4 model in CPRM. Therefore, the anonymous peer review method provides a testing template for further development and can be used to test AI capabilities in other construction project management-related tasks once generative AI models keep improving.

5.1 Theoretical implications

This paper contributed to the literature on AI capabilities in CPRM in several ways. First, the research uncovered the gaps and limitations of existing research, concluding that generative AI and human capabilities have not been comprehensively addressed.

Second, the paper introduced an innovative way that the gap in existing knowledge can be bridged by pitting humans and AI (ChatGPT GPT-4) against each other through a test where a case simulating a real-life CPRM scenario was used to provide data from both AI and human responses. Third, through an anonymous peer review process, construction industry professionals could be used to enable the research to determine if ChatGPT can outperform humans in CPRM.

On this basis, industry stakeholders and researchers can easily continue to develop ways to measure AI capabilities, enabling continuous improvement of CPRM methodologies by leveraging the most scientifically recognized AI capabilities.

5.2 Practical and managerial implications

The practical and managerial implications of this study are substantial, as it highlights the potential for generative AI to enhance decision-making and resource allocation in CPRM. Despite the current limitations of AI technology, particularly in providing context-specific and actionable risk management strategies, it is clear that AI tools like ChatGPT can serve as valuable support to human experts, enhancing efficiency and accuracy in risk management (Hofert, 2023).

The practical implications for practitioners in construction project management include acknowledging the potential of AI, exploring its possibilities, and developing strategies to incorporate AI into their risk management processes. By doing so, practitioners can better leverage AI's strengths and overcome potential barriers and limitations in existing practices, thus creating more robust, efficient, and effective risk management strategies.

While humans have inherent limitations, such as time and motivational constraints that could have potentially influenced their performance, AI models like ChatGPT operate devoid of such restrictions. These disparities suggest that the proposed optimal approach is to leverage AI in CPRM, where its capabilities excel, and freeing human expert resources to concentrate on areas where they demonstrate superior capabilities.

Furthermore, the qualitative results identified the strengths and weaknesses of both ChatGPT and human experts, through the anonymous peer-reviewed test, providing nuanced insights into their respective CPRM capabilities. ChatGPT was praised for its detailed and comprehensive risk identification, analysis, and response strategies, but it was critiqued for partly lacking actionable, context-specific insights.

On the other hand, while ChatGPT excelled in the test, some human experts were able to present more specific, practical, and potentially implementable strategies enforcing the idea that integrating the strengths of AI with human expertise collaboratively could potentially result in the most effective CPRM strategies.

However, there are challenges and limitations of using AI. Both subject matter knowledge and understanding of the application of AI are required. Currently, generative AI (e.g. ChatGPT) can produce false information, and if used poorly it will reap poor results (Ray, 2023, Wach *et al.*, 2023). However, when the AI tools are used well, and provided with sufficient information, they can greatly outperform humans in a variety of tasks (Eloundou *et al.*, 2023), including areas of CPRM.

A proposition for practitioners in CPRM is to acknowledge the potential of AI, incorporate the AI into the CPRM process where it is the most capable, support its weaknesses with human expert collaboration, develop the skills needed in areas such as prompt engineering, and include human experts oversight on generative AI (Barcaui and Monat, 2023). Finally, it could be argued that by creating strategies and capabilities involving the use of AI it is possible to overcome potential barriers and limitations in existing practices.

5.3 Limitations

The research had a number of limitations that should be acknowledged. First, the sample size was limited to sixteen human respondents to the test and nineteen human peer reviewers, additionally, the sample represents Finnish construction industry professionals, which might limit the generalizability of the findings.

Secondly, there were potential variations in the expertise and motivation of the human participants. The simulated nature of the test might have influenced their engagement and performance differently compared to a real-world project scenario.

Third, only the OpenAIs ChatGPT GPT-4 from April 2023 was used as an AI representative, to initially provide one answer for the peer review. Different responses from this model, or the use of alternative AI models, could lead to varied outcomes (OpenAI, 2023) that were covered only via author evaluation by comparison to December 2023 responses from GPT-4.

Fourth, the selected project type and its description were specific and might have produced different responses if a different project type had been chosen. This consideration emphasizes the importance of context and specificity in both AI and human responses.

Fifth, efforts to equalize the base knowledge for the exercise, and the inherent limitations of AI technology (Ray, 2023, Wach *et al.*, 2023) could affect its performance and the outcome of the study.

Finally, the study implemented blind assessments, commonly used in scientific publishing (Shoham and Pitman, 2020) and medical testing (Monaghan *et al.*, 2021) for an objective approach. However, the study nevertheless acknowledges the inherent subjectivity

in human evaluations, which can be influenced by various factors (Kahneman *et al.*, 2021), including the suspicion of responses being AI-generated.

5.4 Future research

For future research, similar research could be conducted in various countries, and on various construction project management process areas and tasks to compare specific capabilities of AI and humans for more comprehensive results. Additionally, over ten thousand AIs for more than two thousand tasks exist (see, e.g. https://theresanaiforthat.com), with more applications of AI emerging constantly, building a comprehensive understanding of their capabilities, and practical applications in the field of CPRM could be considered in the future research.

Furthermore, there is considerable potential to apply AI solutions to project management tasks in practical real-world scenarios to create observations and redesign existing project management processes that consider the efficient use of AI, thus moving beyond the theoretical and controlled conditions of this study.

6. Conclusion

In conclusion, this research advances the theoretical understanding of generative AIs' role in CPRM. It demonstrated that generative AI, as exemplified by the ChatGPT GPT-4, can augment CPRM processes. Quantitatively, ChatGPTs performance, with an average score of 8.6, notably surpasses that of human experts, who scored an average of 5.7. This finding underpins AI's prowess in handling extensive data and crafting intricate responses on CRPM, suggesting a paradigm shift in how risk management is approached in construction project scenarios.

However, the study also contributes to the discourse by highlighting the current limitations of generative AI, represented by OpenAIs ChatGPT GPT-4. The qualitative analysis underscores the importance of context-specific, practical knowledge, and the ability to propose implementable strategies – areas where human experts still hold an edge. ChatGPT responses, while comprehensive, were criticized for their lack of specificity and practicality, highlighting an area for future development. This underscores a pivotal practical implication: the current version of AI tested in CPRM requires a complementary human element to realize its full potential.

Taken together, the findings of this research suggest that the future of CPRM is not a choice between AI and human professionals but rather lies in a synergistic combination of both. The comprehensive and detailed risk assessments provided by AI can serve as a strong basis for risk management, while human professionals, with their practical knowledge and intuition, can further refine and review these assessments and develop practical, context-specific strategies.

This study thus measured ChatGPTs capabilities in CPRM and uncovered the potential where AI can serve as a foundational tool, enhancing human performance and enabling more informed decision-making. Therefore, while AI, in its current state, may not completely surpass human capabilities in CPRM, it undeniably presents a promising tool in the evolution of risk management practices to enhance human performance.

References

Abioye, S.O., Oyedele, L.O., Akanbi, L., Ajayi, A., Davila Delgado, J.M., Bilal, M., Akinade, O.O., Ahmed, A. (2021), "Artificial intelligence in the construction industry: a review of present status, opportunities and future challenges", *Journal of Building Engineering*, Vol. 44, pp. 1-13, doi: 10.1016/j.jobe.2021.103299.

Agarwal, R., Chandrasekaran, S. and Sridhar, M. (2016), *Imagining Construction's Digital Future*, McKinsey & Company, New York.

Akinosho, T.D., Oyedele, L.O., Bilal, M., Ajayi, A.O., Delgado, M.D., Akinade, O.O. and Ahmed, A.A. (2020), "Deep learning in the construction industry: a review of present status and future innovations", *Journal of Building Engineering*, Vol. 101827 No. 32, pp. 1-14, doi: 10.1016/j.jobe.2020.101827.

Aladağ, H. (2023), "Assessing the accuracy of ChatGPT use for risk management in construction projects", *Sustainability*, Vol. 15 No. 16071, pp. 1-27, doi: 10.3390/su152216071.

Arabi, S., Eshtehardian, E. and Shafiei, I. (2022), "Using bayesian networks for selecting risk-response strategies in construction projects", *Journal of Construction Engineering and Management*, Vol. 148 No. 8, pp. 1-19, doi: 10.1061/(ASCE)CO.1943-7862.0002310.

Armstrong, G., Gilge, C., Max, K. and Vora, S. (2023), *Familiar Challenges - New Approaches 2023 Global Construction Survey*, KPMG.

Barcaui, A. and Monat, A. (2023), "Who is better in project planning? Generative artificial intelligence or project managers?", *Project Leadership and Society*, Vol. 4 No. 100101, pp. 1-12, doi: 10.1016/j.plas.2023.100101.

Bolpagni, M., Gavina, R. and Ribeiro, D. (2021), *Industry 4.0 for the Built Environment: Methodologies, Technologies and Skills*, Springer International Publishing, Cham.

Chapman, C. and Stephen, W. (2011), *How to Manage Project Opportunity and Risk: Why Uncertainty Management Can Be a Much Better Approach than Risk Management*, Wiley, Chichester, doi: 10.1002/9781119208587.

Chui, M., Roberts, R., Yee, L., Hazan, E., Singla, A., Smaje, K., Sukharevsky, A. and Zemmel, R. (2023), *The Economic Potential of Generative AI: the Next Productivity Frontier*, McKinsey & Company, New York.

Eloundou, T., Manning, S., Mishkin, P. and Rock, D. (2023), *GPTs Are GPTs: an Early Look at the Labor Market Impact Potential of Large Language Models*, OpenAI, San Francisco.

Ghimire, P., Kim, K. and Acharya, M. (2023), "Generative AI in the construction industry: opportunities and challenges", *arXiv*, pp. 1-30, doi: 10.48550/arXiv2310.04427.

Google (2023), *Gemini: A Family of Highly Capable Multimodal Models*, Google, Mountain View.

Hofert, M. (2023), "Assessing ChatGPT's proficiency in quantitative risk management", *Risks*, Vol. 11 No. 9, pp. 1-37, doi: 10.3390/risks11090166.

International Organization for Standardization (2018), *International Standard ISO 31000:2018 Risk Management – Guidelines*, International Organization for Standardization, Geneva.

Kahneman, D., Sibony, O. and Sunstein, C. (2021), *Noise: A Flaw in Human Judgment*, William Collins, Great Britain.

Kamari, M. and Ham, Y. (2022), "AI-based risk assessment for construction site disaster preparedness through deep learning-based digital twinning", *Automation in Construction*, Vol. 134, 104091, pp. 1-16, doi: 10.1016/j.autcon.2021.104091.

Klepo, M.S., Knežević, D., Knežević, T. and Meštrović, H. (2023), "Artificial intelligence in risk management system on infrastructure projects", *Proceedings of the Creative Construction Conference*, pp. 1-7, doi: 10.3311/CCC2023-028.

Liu, T., Mathrani, A. and Mbachu, J. (2018), "Benefits and barriers in uptake of mobile apps in New Zealand construction industry: what top and middle management perceive", *Facilities*, Vol. 37 Nos 5-6, pp. 254-265, doi: 10.1108/F-08-2017-0078.

Maldonato, M. and Dell'Orco, S. (2011), "How to make decisions in an uncertain world: heuristics, biases, and risk perception", *The Journal of New Paradigm Research*, Vol. 67 No. 8, pp. 569-577, doi: 10.1080/02604027.2011.615591.

Monaghan, T.F., Agudelo, C.W., Rahman, S.N., Wein, A.J., Lazar, J.M., Everaert, K. and Dmochowski, R.R. (2021), "Blinding in clinical trials: seeing the big picture", *Medicina*, Vol. 57 No. 7, pp. 1-13, doi: 10.3390/medicina57070647.

Nyqvist, R., Peltokorpi, A. and Seppänen, O. (2024), "Uncertainty network modeling method for construction risk management", *Construction Management and Economics*, Vol. 42 No. 4, pp. 346-365, doi: 10.1080/01446193.2023.2266760.

OpenAI (2023), *GPT-4 Technical Report*, OpenAI, San Francisco.

Project Management Institute (2016), *Construction Extension to the PMBOK Guide*, Project Management Institute, Newtown Square.

Project Management Institute (2021), *Pulse of the Profession 2021 PMI*, Project Management Institute, Newtown Square.

Rane, N.L. (2023), "Role of ChatGPT and similar generative Artificial Intelligence (AI) in construction industry", *SSRN*, pp. 1-28, doi: 10.2139/ssrn.4598258, available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4598258

Ray, P.P. (2023), "ChatGPT: a comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope", *Internet of Things and Cyber-Physical Systems*, Vol. 3, pp. 121-154, doi: 10.1016/j.iotcps.2023.04.003.

Shoham, N. and Pitman, A. (2020), "Open versus blind peer review: is anonymity better than transparency?", *BJPsych Advances*, Vol. 27 No. 4, pp. 247-254, doi: 10.1192/bja.2020.61.

Smith, N.J., Merna, T. and Jobbling, P. (2006), *Managing Risk: in Construction Projects*, Blackwell Publishing, Oxford.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E. and Lample, G. (2023), *LLaMA: Open and Efficient Foundation Language Models*, Meta AI, New York.

Wach, K., Duong, C., Ejdys, J., Kazlauskaitė, R., Korzynski, P., Mazurek, G., Paliszkiewicz, J. and Ziemba, E. (2023), "The dark side of generative artificial intelligence: a critical analysis of controversies and risks of ChatGPT", *Entrepreneurial Business and Economics Review*, Vol. 11 No. 2, pp. 7-30, doi: 10.15678/EBER.2023.110201.

Zhao, X. (2022), "Evolution of construction risk management research: historiography and keyword co-occurrence analysis", *Engineering, Construction and Architectural Management*, Vol. ahead-of-print No. ahead-of-print, pp. 1-21, doi: 10.1108/ECAM-09-2022-0853.

Corresponding author
Roope Nyqvist can be contacted at: roope.nyqvist@aalto.fi

Appendix

Table A1. The complete quantitative review data on responses

	Human review 1	Human review 2	Human review 3	Human review 4	Human review 5	Human review 6	Human review 7	Human review 8	Human review 9	Human review 10	Human review 11	Human review 12	Human review 13	Human review 14	Human review 15	Human review 16	Human review 17	Human review 18	Human review 19	ChatGPT review
Human response 1																				
Identification														8			7		3	8
Analysis														8			7		3	9
Control														8			6		4	9
Overall														8			7		4	8.7
Human response 2																				
Identification	7						5								5					7
Analysis	4						4								5					8
Control	7						3								8					8.5
Overall	6						4								6					7.8
Human response 3																				
Identification																8				9
Analysis																7				9
Control																8				9
Overall																7.7				9
Human response 4																				
Identification	1					6					6									7
Analysis	2					5					6									7
Control	1					4					6									6
Overall	1.3					5					6									6.5
Human response 5																				
Identification																		8		9
Analysis																		8		6
Control																		7		7
Overall																		8		7.3
Human response 6																				
Identification				5	4				1		5									7
Analysis				5	4				2		5									8
Control				7	4				1		5									7
Overall				6	3				1		5									7.3
Human response 7																				
Identification										9								9		9
Analysis										9								8		9

(*continued*)

Table A1.

	Human review 1	Human review 2	Human review 3	Human review 4	Human review 5	Human review 6	Human review 7	Human review 8	Human review 9	Human review 10	Human review 11	Human review 12	Human review 13	Human review 14	Human review 15	Human review 16	Human review 17	Human review 18	Human review 19	ChatGPT review
Control										9								9		8
Overall										8								9		8.7
Human response 8																				
Identification					5							6	3	3			5			7
Analysis					5							3	0	1			4			5
Control					5							4	3	2			6			7
Overall					5							4	2	2			5			6.3
Human response 9																				
Identification																			7	8
Analysis																			7	8
Control																			7	7
Overall																			7.5	7.7
Human response 10																				
Identification																				6
Analysis																				5
Control																				6
Overall																				5.7
Human response 11																				
Identification								9												8
Analysis								7												7
Control								7												8
Overall								7												7.7
Human response 12																				
Identification		7														7				9
Analysis		7														7				9
Control		9														8				9
Overall		7														7.3				9
Human response 13																				
Identification							10	6	8											9
Analysis							7	5	2											8
Control							7	6	4											7
Overall							8	5	5											8

(*continued*)

Table A1.

	Human review 1	Human review 2	Human review 3	Human review 4	Human review 5	Human review 6	Human review 7	Human review 8	Human review 9	Human review 10	Human review 11	Human review 12	Human review 13	Human review 14	Human review 15	Human review 16	Human review 17	Human review 18	Human review 19	ChatGPT review
Human response 14																				
Identification		9				9				8										8
Analysis		8				8				7										6
Control		7				8				4										4
Overall		8				8				7										6
Human response 15																				
Identification				4								5	5							7
Analysis				5								4	3							7
Control				4								6	5							8
Overall				4								5	4.3							7.3
Human response 16																				
Identification															7					9
Analysis															7					8
Control															9					8
Overall															7.5					8.3
ChatGPT response																				
Identification	9	10		8	10	10	8	6	6	9	7	9	9	9	10	9	8	9	2	10
Analysis	9	10		9	10	10	10	9	9	7	7	9	9	6	9	9	8	9	3	9
Control	9	10		9	10	9	10	9	10	8	7	10	9	6	9	9	9	8	4	8
Overall	9	10		9	10	10	9.3	8	9	9	7	10	9	7	9	9	8	9	3	9

Note(s): *Provided only qualitative assessment