Journal of Healthcare Simulation

Check for updates

Journal of Healthcare Simulation

ORIGINAL RESEARCH

Can an experienced qualitative researcher distinguish AI from human qualitative content analysis?

Alexandra T. Lucas[1], Jianna Ramos[2], Maria Bajwa[3], Aaron Calhoun[4], Mark W. Scerbo[5], Janice C. Palaganas[3]

[1]*Department of Pediatrics, MassGeneral Brigham for Children, Slavin Academy, Massachusetts General Hospital and Mass General Brigham for Children, Boston, Massachusetts, USA*
[2]*Department of Biochemistry, Case Western Reserve University, Cleveland, Ohio, USA*
[3]*MGH Institute of Health Professions, Boston, Massachusetts, USA*
[4]*Department of Pediatrics, University of Louisville School of Medicine and Norton Children's Medical Group, Louisville, Kentucky, USA*
[5]*Department of Psychology, Old Dominion University, Norfolk, Virginia, USA*

Corresponding author: Alexandra T. Lucas, atlucas@mgb.org

https://johs.org.uk/article/doi/10.54531/FCQK4553

ABSTRACT

Background

Artificial intelligence (AI) has become increasingly embedded in research workflows. Large language models (LLMs) are being used to code segments of text, organise codes into themes and interpret patterns within contexts. Recent comparisons between human and AI analyses demonstrate up to 80% thematic overlap, yet humans consistently exhibit deeper interpretive integration and contextual understanding. This study assesses whether experienced researchers can distinguish between entirely human-generated and AI-generated qualitative content analyses of a simulation debriefing.

Methods

We conducted a qualitative descriptive study comparing human-generated qualitative content analysis (QCA) with ChatGPT-4o-generated QCA using a single focus group transcript on emotion management during debriefing of an emotion-triggering simulation case. First, a human research team performed QCA following Schreier's framework to identify themes. The same transcript was then analysed by ChatGPT-4o using a prompt aligned with Schreier's QCA approach. Second, both outputs were standardised into matched text formats to control for structural and stylistic variation. Third, de-identified versions of each analysis (human and AI) were provided to a separate group of experienced reviewers, who completed two rounds of open-ended review. Finally, we conducted a QCA of the reviewers' responses to examine how expert researchers characterise and differentiate AI- and human-generated qualitative analyses.

Results

Only three out of five experienced researchers were able to distinguish between AI-generated and human-generated QCA. The reviewers who correctly identified the AI analysis noted that AI generated incomplete quotes and missed specific details. Reviewers' preferences for a given analysis did not correlate with accurate identification of human-generated vs AI-generated analysis. Three out of five reviewers preferred the AI analysis.

Submission Date: 29 March 2026
Accepted Date: 22 June 2026
Published Date: 10 August 2026

Discussion
To our knowledge this is the first study specific to simulation debriefing. Two out of five reviewers were not able to differentiate between human and AI-generated outputs, suggesting that confident identification of AI-generated content may reflect bias more than accuracy, and raising concerns about the fairness of informal AI-detection practices. A combined approach in which humans perform an initial analysis that is then followed by subsequent refinement using AI-based processes may represent the most productive path forward.

Introduction

As artificial intelligence (AI) becomes increasingly embedded in research workflows, a critical question emerges: can experienced researchers distinguish between AI-generated and human-generated analysis, and how do they do so? This ability has direct implications for research integrity, peer review, methodological transparency and the trustworthiness of published findings. If AI-generated analysis is indistinguishable from human analysis, reviewers and consumers of research may unknowingly accept outputs that lack the contextual reasoning central to qualitative inquiry.

AI, particularly generative AI, has rapidly expanded to perform tasks associated with human cognition, including language interpretation, synthesis and analysis. Large language models (LLMs) are increasingly used in literature review and quantitative analysis [1]. In qualitative research, analysis involves coding segments of text, organising codes into themes and interpreting patterns within a context [2]. LLMs can be trained to replicate this process [3], producing outputs that may resemble human work.

Recent comparisons between human and AI analyses demonstrate up to 80% thematic overlap. Humans consistently exhibit deeper interpretive integration and contextual understanding [3]. However, AI cannot yet replicate the critical judgement central to qualitative inquiry [1], and errors persist [4,5]. Misattributing AI-generated analysis as human-produced risks undermining the epistemological foundations of qualitative research.

Qualitative content analysis (QCA) of simulation debriefings poses a unique challenge for AI, as these are reflective learning processes guided by a skilled facilitator, resulting in interactional dynamics that must be considered. Because debriefings offer a rich window into student learning, emotional processing and professional development, analysis of the debriefings can be particularly useful for educators. Specific frameworks have been developed to help codify debriefings [6], and there is ongoing research to understand how debriefing works and how it enables learning [7]. This means that QCA will likely be an oft-used method to both understand individual debriefings as well as to deepen our understanding of debriefing as a practice.

This study assesses whether experienced researchers can distinguish between entirely human-generated and AI-generated qualitative content analyses.

Methods

We conducted a qualitative descriptive study comparing human-generated QCA with ChatGPT-4o-generated QCA using a single focus group transcript on emotion management during debriefing of an emotion-triggering simulation case (SDC1). First, a human research team performed QCA following Schreier's framework [2] to identify themes. The same transcript was then analysed by ChatGPT-4o using a prompt aligned with Schreier's QCA approach. Second, both outputs were standardised by the authors who did not participate in the QCA into matched text formats to control for structural and stylistic variation (Table 1). Third, de-identified versions of each analysis (human and AI) were provided to a separate group of experienced reviewers, who completed two rounds of open-ended review with specific prompts (SDC2). Reviewers were chosen based on their prior experience in qualitative research. In Round 1, reviewers independently compared thematic similarities, differences, depth of insight and overall preference. In Round 2, they identified which analysis they believed was AI-generated and reflected on their personal potential biases towards AI. Finally, we conducted a QCA of the reviewers' responses to examine how expert researchers characterise and differentiate AI- and human-generated qualitative analyses (Figure 1).

This study was deemed exempt by the Mass General Brigham Institutional Review Board.

Results

Only three out of five experienced researchers were able to distinguish between AI-generated and human-generated QCA (Table 2). The reviewers who correctly identified the AI-analysis noted that AI generated incomplete quotes and missed specific details (such as identifying positive and negative emotions, as opposed to just negative ones). Those who incorrectly identified the AI-analysis thought that the AI analysis 'provided more insight' as compared with the human analysis.

Interestingly, the reviewers' preferences for a given analysis (human vs AI) did not correlate with accurate identification of human-generated vs AI-generated analysis. Three out of five reviewers actually preferred the AI analysis, citing that it was 'more useful as debriefer' and had 'greater richness and insight'. Among those who preferred the AI analysis, only one identified it as their preferred analysis. One reviewer believed that neither analysis was sufficient.

Overall, comparisons of depth, scope and insight revealed that the human-generated analyses contained broad and simplistic language, whereas the AI-generated responses used detailed language that was organised by time. Reviewers noted that the human-generated analysis identified a potentially important theme that was not addressed in the AI analysis.

Table 1: Qualitative content analyses, generated by human research team and AI

Human research team – key themes identified	AI – key themes identified
A. Individual emotions Participants expressed strong individual emotional reactions to simulation events, including: ● **Positive** – from feeling supported, not feeling rushed and feeling comfortable. ● **Negative** – primarily from lingering anxiety, frustration and disappointment from case performance. ● **As a result of triggers** – events that occurred in the case. ● **As a result of debriefing** – debriefing allows the processing of individual emotions. 'I was frustrated with myself because as a prior urinary nurse, I felt like I should have known what to do… I just felt frustrated with myself and kind of incompetent'. 'For me, it was how he made me very comfortable… there was no rush and covered so much, allowed us to process… I never got a feeling that I was judged during that debrief'.	A. Emotional responses to simulation Participants expressed strong emotional reactions to simulation events, including: ● **Stress and anxiety** – from time pressure, unexpected DNR presentation and confusing instructions. ● **Frustration and disappointment** – primarily directed at self-performance, perceived knowledge gaps or not meeting team expectations. ● **Inadequacy** – due to time away from clinical practice or missed critical details. 'I felt frustrated with myself… I haven't been in that kind of environment for 20 years'. 'We were all trying to work as a team… we were frustrated with our individual performance'.
B. Team emotions The emotional state of the team: ● **Positive** – being part of a team and not alone. ● **Negative** – primarily from lingering anxiety, frustration and disappointment from case performance. 'I felt there was frustration that we were all trying to work as a team. We weren't frustrated with each other. We were frustrated with our individual performance as part of a team practice'.	B. Emotional responses during debriefing The emotional state carried into the debriefing, influenced by: ● **Processing missed elements** (e.g. DNR, medication errors) ● **Team reflection** on roles and communication ● **Self-awareness** of limitations or errors 'We were processing our teamwork… then he threw in, did you look at the IV?'
C. Group culture There were group norms and expectations 'A few ground rules, no right or wrong answers. One person speaks at a time because we are recording. And, again, with confidentiality… whatever happens or stays or we do the Vegas rule in all the courses as well'.	C. Effective debriefing strategies Facilitators' approaches that supported emotional processing included: ● **Patient, non-judgemental listening** ● **Open-ended questioning** ● **Validating and reflecting participant emotions** ● **Structured inquiry and evidence-based reflection** ● **Allowing participants to process at their own pace** 'He was a patient listener'. 'He validated all our responses. I felt comfortable'. 'He allowed us to process… didn't rush'.
D. Debriefer strategies Strategies were determined to be effective and ineffective. ● **Effective strategies** included: ○ Communication techniques – clarification and understanding, questioning strategies, rephrasing and reflecting ○ Emotional support and validation ○ Facilitation and debriefing approaches – structured approach, facilitating the conversation, allowing emotional and cognitive processing ○ Listening and presence ○ Team engagement and inclusivity	D. Ineffective or mixed strategies Some feedback indicated opportunities for improvement: ● Introducing **multiple surprises too quickly** without allowing time to process the first ● Inconsistent focus between **team dynamics vs clinical content**, depending on the facilitator 'Maybe allow us to sink in with the first surprise before going to the next'. 'He went to the medical piece, and you didn't include that… it created some differences'.

Table 1. Continued

Human research team – key themes identified	AI – key themes identified
● **Ineffective strategies** included: ○ Surprise elements and impact – these involve disclosing unexpected twists in the case or debrief that can affect learner processing and reflection. ○ Facilitator debriefing approaches that do not allow team reflection and learning. ○ Information sharing strategies – not using best judgement on how, when and what kind of information was provided during the debrief or simulation. ○ Cultural and team dynamics – not recognising or understanding differences within the team that influence learning and interaction. ○ Emotional processing – missing emotional cues during debriefing. ○ Missing opportunities or approaches that hindered the learning process. 'Maybe allow us to sink in with the first surprise before going to the next'. 'He went to the medical piece, and you didn't include that… it created some differences'. 'He was a patient listener… He patiently listened while we processed and I thought that was, that's very safe'.	E. Emotional safety and learning environment Participants emphasised that a **safe, respectful and non-judgemental environment** was crucial for open reflection: ● Reduced anxiety ● Encouraged vulnerability ● Promoted deeper learning 'It was very comfortable… I never got a feeling that I was judged'. 'We were already judging ourselves… but he was not judging us'. F. Takeaways and best practices Participants recommended several best practices for emotionally charged debriefings: ● **Patient, empathetic listening** ● **Individual check-ins** ● **Time for emotional and cognitive processing** ● **Clarify learning objectives in advance** ● **Balance between clinical and team performance feedback** 'Clarify the medical facts first, so that everyone's open to listening to what's next'. 'Validate the emotions and reflect what you think you're hearing'.
E. Debriefer persona Participants emphasised that a **direct, authentic, patient and non-judgemental environment** was crucial for open reflection. 'It was very comfortable… I never got a feeling that I was judged'. 'We were already judging ourselves… but he was not judging us'.	

Reviewer self-reflections on AI use in analysis generally supported ethical AI use, and none of the reviewers supported either fully human or fully AI analysis, despite several being unable to distinguish AI from human output. Some felt that AI would be best served to organise and proofread, whereas others felt that AI could be used in any aspect of analysis as long as humans provide oversight and control over each step. One reviewer likened AI use to driving a car: while the vehicle can get a driver to a destination more quickly, the driver needs to know the vehicle's limitations and maintain control.

Discussion

We found that two out of five researchers could not differentiate between human and ChatGPT-4o outputs. Despite reliance on a single LLM and a small sample of reviewers without use of AI-enhanced qualitative software (i.e. ATLAS.ti or NVivo), this finding carries significant implications beyond the research context. Across academia, educators and reviewers are increasingly flagging work they believe to be 'AI-generated', yet this study suggests that human judgement alone is a poor standard for making that determination, an observation reinforced in other recent studies [8]. Indeed, our correct detection rate of AI-generated output (60%) is even higher than prior studies, where humans only had an accuracy of 10% for correctly detecting AI-generated text [8]. Confident identification of AI-generated content may reflect bias more than accuracy, raising concerns about the fairness of informal AI-detection practices.

Although other studies have evaluated the role of AI in QCA, to our knowledge this is the first study specific to simulation debriefing. Transcripts from debriefing represent a unique dataset, as they are reflective learning processes guided by a skilled facilitator. Because the facilitator and participants can influence one another, interactional dynamics must be considered more heavily than in a traditional interview. Debriefings often follow a particular educational theory that can be used to guide analysis, which an AI model may not be aware of and therefore may provide a less useful coding structure. Additionally, debriefings may be emotionally charged sessions, as in our study. AI has previously struggled with detecting emotional nuances, such as emotional shifts where the emotion changes from one utterance to another. Our findings agree with this, AI was only able to identify negative emotions, whereas the humans identified positive and negative valences. Additionally, cultural variations in emotional expressions are another challenge for AI, particularly given the diversity of those participating in simulations [9,10]. While AI's emotional detection is improving, this is an important limitation to its ability in analysing debriefing transcripts [11]. Given how difficult it is for human reviewers to be able to accurately determine what analysis has been written by an LLM and

Figure 1: Study design

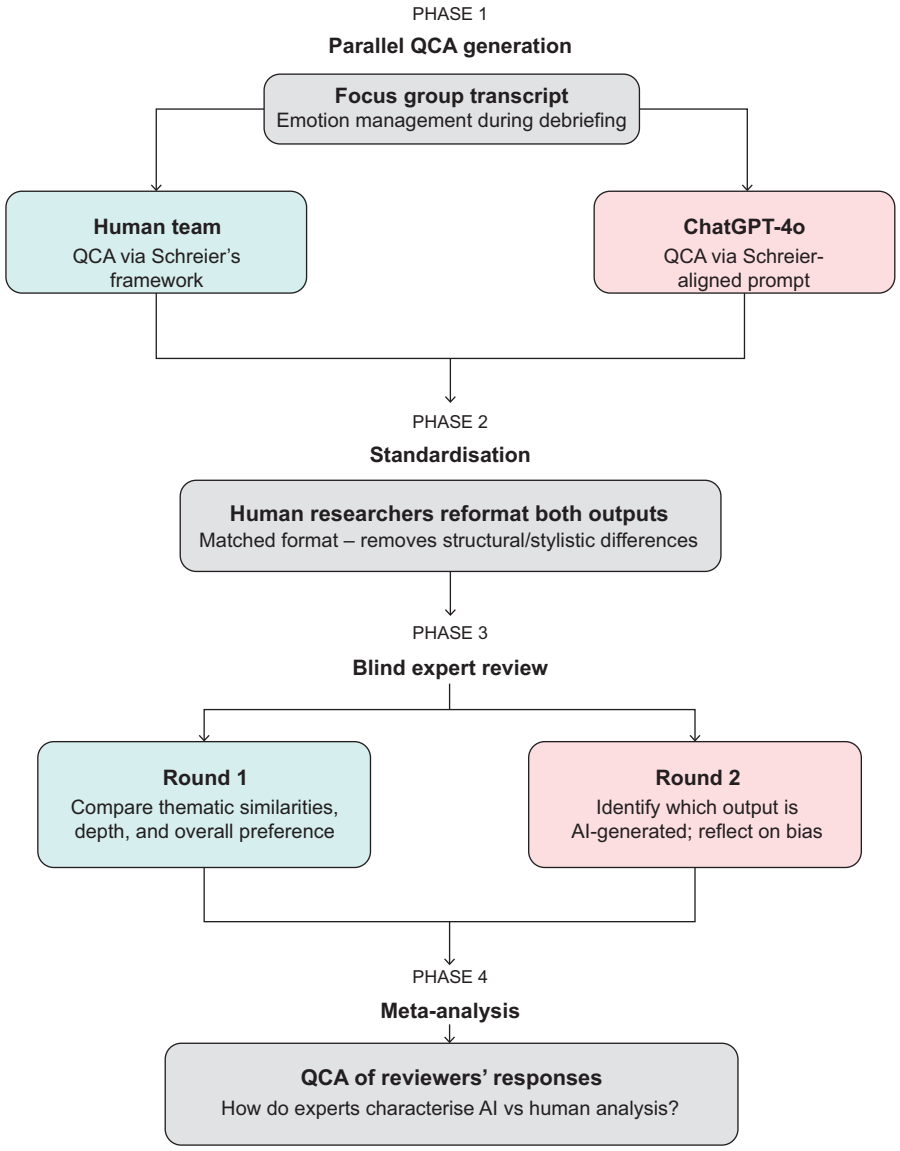

Table 2: Expert human review of AI- and research team-generated QCA output

Domain	Human research team-generated	AI-generated
Overall thematic structure	Themes organised more globally, without explicit temporal segmentation.	Themes organised temporally across phases of the simulation (pre-briefing, simulation, debriefing).
Thematic depth and specificity	Described as 'simplistic', 'non-specific' and 'bland'.	Described as 'richer', 'more focused' and 'more descriptive'.
Unique contributions	Identified 'group culture' as a potentially important theme absent from the AI analysis.	Did not identify group culture as a distinct theme.
Insight into debriefing strategies	Two reviewers felt this analysis provided 'more detail about effective and ineffective debriefing'.	Two reviewers found this analysis 'more specific' and 'richer'.
Reviewer preference	Preferred by one reviewer	Preferred by three of five reviewers, citing 'more useful categories'.
Identification of AI vs human	Incorrectly identified as AI-generated by two reviewers due to 'buzzwordy language'.	Correctly identified as AI-generated by three reviewers due to conciseness, methodological adherence and levels of bulleting.

what has been written by a human, our findings suggest that analyses must have human oversight in order to avoid missing these important nuances.

Similar to prior studies, we found high thematic overlap between AI and human analyses [5], suggesting GPTs reliably identify themes within a dataset. Although the human analysis was longer than the AI, they identified a similar number of themes. This overlap is what makes detection difficult and why assumptions about what AI-generated work 'looks like' can be unreliable. Previous studies using LMs in QCA have similarly found that GPT-4 may fail to identify low-incidence but important concepts [12].

Researchers disagreed on which analysis they preferred, with some favouring the organisation of the AI output while others preferred the insight of the human approach. This reflects the unique interpretive lens each researcher brings, shaped by expertise and positionality. While researchers can document reasoning through reflexive notes, LLMs do not generate output via similar decision-making, making analytic choices harder to probe.

Notably, the human analysis identified a potentially important theme that AI missed, highlighting the continued necessity of human judgement in qualitative inquiry. A combined approach in which humans perform an initial analysis that is then followed by subsequent refinement using AI-based processes may represent the most productive path forward. This ensures that potential AI hallucinations and biases do not embed themselves in the initial framework or shape human perception of the data in a pre-critical fashion, a point supported in recent work on the ethics of AI in scholarship [13].

As AI evolves, it raises ethical concerns around privacy and data protection, signifying the need for ongoing evaluation of AI-generated qualitative outputs and clear standards for transparency and attribution in research.

Supplementary material

Supplementary data are available at *Journal of Healthcare Simulation* online.

Acknowledgements

The authors would like to thank Anne Weaver, Jared Kutzin, Debra Nestel, Suzie Kardong-Edgren, Cathy Deckers, Traci Grove, Lulu Sherif Mahmood and MaryAnn Martin for their contributions to this work through the MGH IHP Healthcare Simulation Research Course.

Declarations

Authors' contributions

ATL and JCP participated in the conceptualisation, planning and design of the process described in this article. All authors conducted the research and data collection. ATL, MB and JCP conducted the analysis. All authors contributed to the writing of the manuscript. All authors have followed the instructions for authors and have read and approved the manuscript.

Funding

There were no sources of funding for this study.

Availability of data and materials

None declared.

Ethics approval and consent to participate

None declared.

Competing interests

There are no conflicts of interest to disclose at this time.

References

1. Christou PA. How to use thematic analysis in qualitative research. Journal of Qualitative Research in Tourism. 2023;1:1–17. doi: 10.4337/jqrt.2023.0006
2. Schreier M. Qualitative content analysis in practice. London, UK: Sage Publications Ltd. 2012.
3. Morgan DL. Exploring the use of artificial intelligence for qualitative data analysis: the case of ChatGPT. International Journal of Qualitative Methods. 2023;22. doi: 10.1177/16094069231211248
4. Cheligeer C, Yang L, Nandi T, Doktorchik C, Quan H, Zeng Y, et al. Natural language processing (NLP) aided qualitative method in health research. Journal of Integrated Design and Process Science. 2024:27(1)41–58. doi: 10.3233/JID-220013
5. Bijker R, Merkouris SS, Dowling NA, Rodda SN. ChatGPT for automated qualitative research: content analysis. Journal of Medical Internet Research. 2024;26:e59050. doi: 10.2196/59050
6. Berger-Estilita J, Lüthi V, Greif R, Abegglen S. Communication content during debriefing in simulation-based medical education: an analytic framework and mixed-methods analysis. Medical Teacher. 2021;43(12):1381–1390. doi: 10.1080/0142159X.2021.1948521
7. Ranjev K, Reedy G. Transforming professional identity in simulation debriefing: a systematic metaethnographic synthesis of the simulation literature. Simulation in Healthcare: The Journal of the Society for Simulation in Healthcare. 2024;19(2):90–104. doi: 10.1097/SIH.0000000000000734
8. Cheng A, Lin Y, Reedy G, Joseph C, Wirkowski S, Mallette V, et al. Ability of AI detection tools and humans to accurately identify different forms of AI-generated written content. Advances in Simulation 2025;10(1):66. doi: 10.1186/s41077-025-00396-6
9. Elyoseph Z, Refoua E, Asraf K, Lvovsky M, Shimoni Y, Hadar-Shoval D. Capacity of generative AI to interpret human emotions from visual and textual data: pilot evaluation study. JMIR Mental Health. 2024; 11:e54369. doi: 10.2196/54369
10. Grove TR, Lucas AT, Martin M, et al. Am I as effective at identifying emotions as artificial intelligence? A comparative study of emotion recognition. Cureus Journal of Computer Science2025;2:es44389-025-07224-y. doi: 10.7759/s44389-025-07224-y
11. Poria S, Majumder N, Mihalcea R, Hovy E. Emotion recognition in conversation: research challenges, datasets and recent advances. 2019 arXiv:1905.02947.
12. Salazar M, Chaw M, Hellier Y, Hsia S, Gruenberg K. Comparison of qualitative analyses conducted by artificial intelligence versus traditional methods. AJPE. 2025;89(12).
13. Cheng A, Calhoun A, Reedy G. Artificial intelligence-assisted academic writing: recommendations for ethical use. Advances in Simulation. 2025;10(1):22.