Original Paper

Use of Large Language Models for Structured Quantitative and Qualitative Analysis of Simulation Debriefings Using the DASH Framework (AIDE Study): Prospective Observational Study

Julien Pico[1,2], MD; Tarikeddine Bouhadiba[1], MD; Shiva Jazestani[1], MD; Eric Zhu[1], MD; Ba-Khang Nguyen[1]; Reda Baroudi[1], MD; Said Dababneh[1], MD; Issam Tanoubi[1,3], MD, PhD

[1]Centre d'apprentissage des attitudes et habiletés cliniques, Université de Montréal, Montréal, QC, Canada

[2]Department of Maternal, Child and Women's Anesthesiology and Intensive Care Medicine, Pediatric Anesthesia Unit, Montpellier University Hospital, Montpellier, France

[3]Centre de Recherche en Pédagogie de la Santé, Faculté de Médecine, Université de Montréal, Montréal, QC, Canada

Corresponding Author:

Julien Pico, MD
Centre d'apprentissage des attitudes et habiletés cliniques
Université de Montréal
Pavillon Roger-Gaudry, 2900, boul. Édouard-Montpetit, 8e étage, local N-805
Montréal, QC H3T 1J4
Canada
Phone: 1 625795440
Email: j-pico@chu-montpellier.fr

Abstract

Background: Simulation-based education is a cornerstone of health care training, with debriefing representing the critical phase of a learning process. However, assessing debriefing quality remains challenging, relying largely on expert judgment and structured tools such as the Debriefing Assessment for Simulation in Healthcare (DASH). The emergence of AI, particularly large language models (LLMs), offers new opportunities for the automated analysis of educational interactions.

Objective: This study aimed to explore the feasibility and characteristics of AI-automated analysis of postsimulation debriefings using the DASH framework.

Methods: We conducted a prospective, single-center, observational study at the *Centre d'Apprentissage des Attitudes et Habiletés Cliniques* (*Université de Montréal*). Postsimulation debriefings involving anesthesia residents in postgraduate years 2 and 4 were conducted by a single instructor, audio-recorded, transcribed, and anonymized. Transcripts were analyzed using a natural language processing model (ChatGPT) with a structured prompt aligned with the 6 DASH elements. For each debriefing, the model generated a structured qualitative evaluation, including a rating for each DASH element using the original 7-point DASH effectiveness scale and an identification of strengths and areas for improvement, supported by transcript excerpts. No comparison with a human DASH-trained evaluator was performed.

Results: A total of 30 debriefings were included, with no exclusions or technical issues. A complete, structured AI-generated evaluation across all 6 DASH elements was obtained for all 30 (100%) debriefings, meeting the predefined feasibility outcome. "Establishes an engaging learning environment," "Maintains an engaging learning environment," "Provokes engaging discussion," and "Identifies and explores performance gaps" each had a median rating of 6 (IQR 6-7, 6-6, 6-7, and 6-6, respectively). "Structures the debriefing in an organized way" had the lowest median rating, at 4 (IQR 4-5), while "Helps trainees achieve or sustain good future performance" had a median rating of 5 (IQR 5-6). A cross-debriefing analysis identified recurring patterns across sessions, including consistent learner engagement and opportunities to improve debriefing structure and the formulation of strategies to support future performance.

Conclusions: This study demonstrates the feasibility of AI-assisted qualitative assessment of postsimulation debriefings using an LLM prompted in accordance with the DASH framework. AI may represent a complementary tool to support pedagogical analysis, instructor reflection, and faculty development in simulation-based education. Although this exploratory study focused on textual transcript analysis without incorporating nonverbal communication or expert benchmarking, the findings support

the feasibility of AI-assisted debriefing assessment and provide a foundation for future validation studies. Further research comparing AI-generated and expert evaluations could help refine these approaches and inform their integration into simulation-based education and faculty development.

JMIR Form Res 2026;10:e101797; doi: 10.2196/101797

Keywords: artificial intelligence ; AI; large language models ; simulation-based education ; debriefing ; Debriefing Assessment for Simulation in Healthcare framework; DASH framework ; medical education

Introduction

Simulation-based training has become a major component of the education of health care professionals. It supports the acquisition and refinement of both technical and nontechnical skills in a safe learning environment and contributes to improved clinical performance and patient safety [1]. Within this educational approach, debriefing represents a critical phase of the learning process, promoting reflective practice, knowledge integration, and the development of clinical reasoning [2]. A systematic review demonstrated that debriefing significantly improves learners' competencies, particularly in communication, teamwork, and clinical performance [3].

Despite its central educational role, assessing debriefing quality remains challenging. Evaluation often relies on instructor or expert judgment and may be associated with substantial interrater variability [4-6]. Structured frameworks have therefore been developed to guide debriefing practices and support reflective practice [7]. Among these, the Debriefing Assessment for Simulation in Healthcare (DASH) is one of the most widely used instruments for evaluating debriefing quality across its 6 core elements [8]. However, psychometric studies have reported variability in reliability depending on the educational context and evaluator training [9,10]. Furthermore, structured debriefing assessment requires specific pedagogical expertise and considerable time investment, which may limit its routine implementation in clinical simulation programs [10].

AI, particularly natural language processing approaches based on large language models (LLMs), is rapidly expanding in medical education. Recent reviews highlight emerging applications in feedback generation, performance assessment, adaptive learning, and educational analytics [11-14]. Within simulation-based education, exploratory studies have begun to investigate the use of AI for debriefing analysis, suggesting its potential to identify reflective markers, analyze communication patterns, and generate structured feedback [15-17]. Other studies have reported favorable acceptability among learners and educators [18,19], as well as potential applications for faculty development and the continuous improvement of teaching practices [20,21].

Despite this growing interest, empirical evidence remains limited regarding the feasibility of LLM-generated debriefing analysis in authentic educational settings and the nature of the evaluations produced using a validated educational framework such as DASH. Few studies have systematically described the qualitative profiles generated by AI from real-world debriefing sessions.

The Artificial Intelligence for Debriefing Evaluation (AIDE) study was designed within this exploratory framework. The objective of this study was to describe the feasibility and characteristics of AI-generated qualitative analysis of postsimulation debriefings using an LLM prompted according to the original DASH framework.

Methods

Study Design and Setting

This prospective, single-center, observational study aimed to describe the feasibility and characteristics of automated quantitative and qualitative analysis of postsimulation debriefings using an LLM structured according to the DASH framework. Comparison with human DASH-trained raters was outside the scope of this initial feasibility study, which was designed to characterize the ability of the LLM to generate complete and structured DASH-aligned evaluations.

The study was conducted at the *Centre d'Apprentissage des Attitudes et Habiletés Cliniques* of the *Université de Montréal*. Simulation trainees were anesthesia residents (postgraduate years 2 and 4) participating in high-fidelity simulation sessions followed by structured debriefings facilitated by an instructor with more than 10 years of experience in simulation-based education. Multiple high-fidelity simulation scenarios were used during the study period. Scenarios were not assigned according to postgraduate year, and residents from different postgraduate years could participate in the same simulation session. To limit variability related to differences in facilitation style, pacing, questioning strategies, and debriefing structure across instructors, all analyzed debriefings were conducted by the same instructor. Debriefings were consecutively included between January 22, 2026, and April 23, 2026.

Ethical Considerations

The study was approved by the research ethics committee in *Comité d'éthique de la recherche en éducation et en psychologie* (CEREP), *Université de Montréal* (2025-7463). The instructor conducting the debriefings provided informed consent for audio recording and anonymized analysis of the debriefings. Data were stored on secure institutional servers in accordance with applicable regulations. No financial or nonfinancial compensation was provided to the simulation trainees or the instructor for their participation in this study.

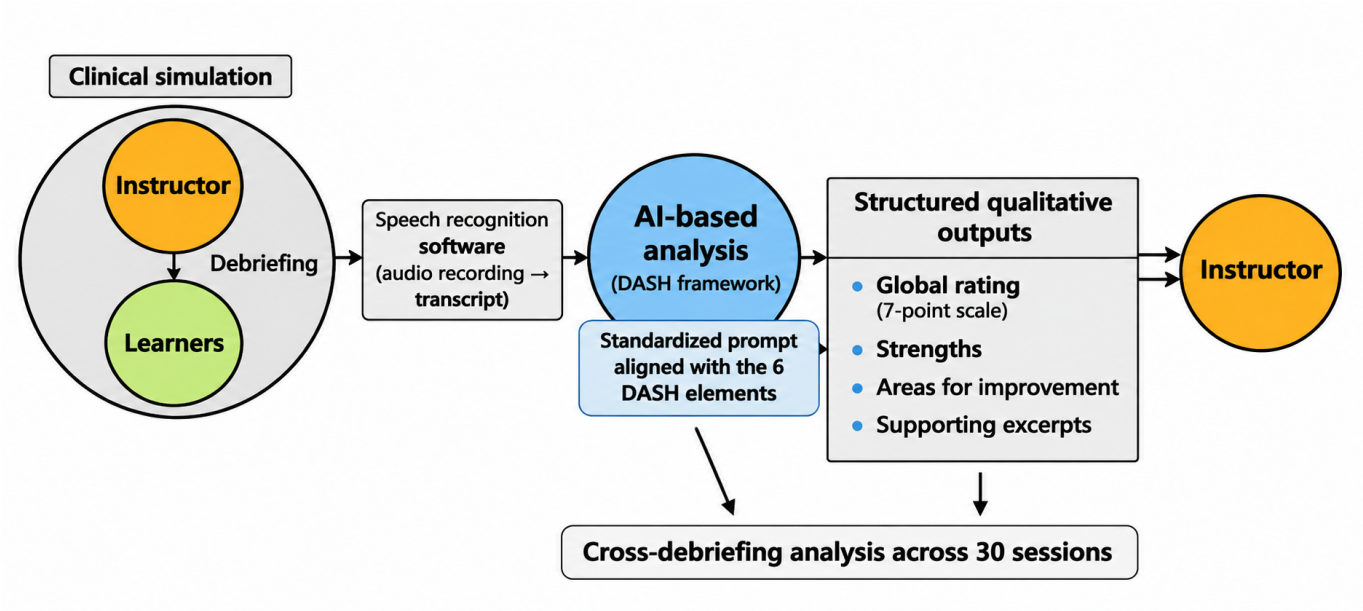
Data Collection and Transcription

Each debriefing session was audio-recorded using a secure institutional device. Recordings were transcribed and

anonymized prior to analysis, and all identifiable personal or contextual information was removed. Audio recordings were checked for technical quality and audibility, and transcripts were reviewed for overall coherence before AI analysis. No systematic utterance-by-utterance verification against the complete audio recordings or dedicated speaker-diarization procedure was performed. Only debriefings with complete and technically usable recordings were included.

Following completion of the study period, a cross-debriefing analysis of all 30 sessions was performed to identify recurring patterns. A single aggregate feedback report was provided to the instructor at the end of the study rather than after each individual debriefing (Figure 1).

Figure 1. Workflow of the AI-assisted debriefing analysis process in the Artificial Intelligence for Debriefing Evaluation (AIDE) study. DASH: Debriefing Assessment for Simulation in Healthcare.



Debriefing sessions conducted during clinical simulation were audio-recorded and transcribed using speech recognition software. Transcripts were analyzed using an AI model structured according to the DASH framework, which generates structured qualitative outputs (global rating, strengths, areas for improvement, and supporting excerpts) for each individual debriefing. Following completion of the study period, a cross-debriefing analysis (N=30) was performed to identify patterns across sessions, and a single, aggregate feedback report was provided to the instructor at the end of the study, rather than after each debriefing.

AI-Based Debriefing Analysis

Transcripts were analyzed using the LLM ChatGPT (GPT-5.3; OpenAI). Transcripts were submitted individually through the ChatGPT web interface using the same standardized prompt and reference materials for each analysis. No API or custom model parameters were used. Before the analyses, the model was provided with the original DASH publication, the official DASH score sheet, and several peer-reviewed articles applying the DASH framework as reference materials. The *DASH Rater's Handbook* was not provided. A structured prompt was specifically designed to align model outputs with the 6 official DASH elements and their corresponding dimensions and behaviors: "Establishes an engaging learning environment," "Maintains an engaging learning environment," "Structures the debriefing in an organized way," "Provokes engaging discussion," "Identifies and explores performance gaps," and "Helps trainees achieve or sustain good future performance." The complete standardized prompt used for the analyses is provided in Multimedia Appendix 1. For each element, the model generated a rating using the original 7-point DASH effectiveness scale. In addition to these ratings, the model was prompted to generate an in-depth qualitative analysis for each debriefing, identifying 3 principal strengths and 3 main areas for improvement, supported by illustrative excerpts from the transcript. The model was explicitly instructed to base its analysis exclusively on the provided transcript and to use only supporting excerpts contained within the transcript, without introducing information not present in the source material.

The analysis was conducted in 2 steps. First, each debriefing was analyzed individually to generate a detailed evaluation and a rating for each of the 6 DASH elements using the original 7-point DASH effectiveness scale. Second, a cross-debriefing comparative analysis was performed by examining AI-generated evaluations across all sessions to identify recurring patterns, element-specific trends, areas of consistency or variability, and common strengths and limitations in debriefing practices. This comparative approach enabled the generation of an overall qualitative profile of debriefing practices within the dataset.

The prompt was iteratively refined during pilot testing to standardize terminology, the level of detail across DASH elements, the number of strengths and areas for improvement,

the use of supporting transcript excerpts, and the formatting of ratings. These refinements focused on output structure and consistency rather than on modifying the DASH evaluation criteria. The final prompt was then applied consistently to all analyzed debriefings, and model-generated results were not subsequently modified.

Outcomes and Statistical Analysis

The primary outcome was the feasibility of AI-based debriefing analysis, operationalized as the proportion of debriefings for which the model generated a complete, structured evaluation across all 6 DASH elements. A complete evaluation was defined as including a rating using the original 7-point DASH effectiveness scale for each element, 3 strengths, 3 areas for improvement, and supporting transcript excerpts. Secondary outcomes included the distribution of ratings across the 6 DASH elements and the identification of recurring strengths and areas for improvement across debriefings. Given the exploratory nature of the study, analyses were descriptive, without inferential statistics. DASH ratings are reported as median values with IQRs for each of the 6 elements.

Results

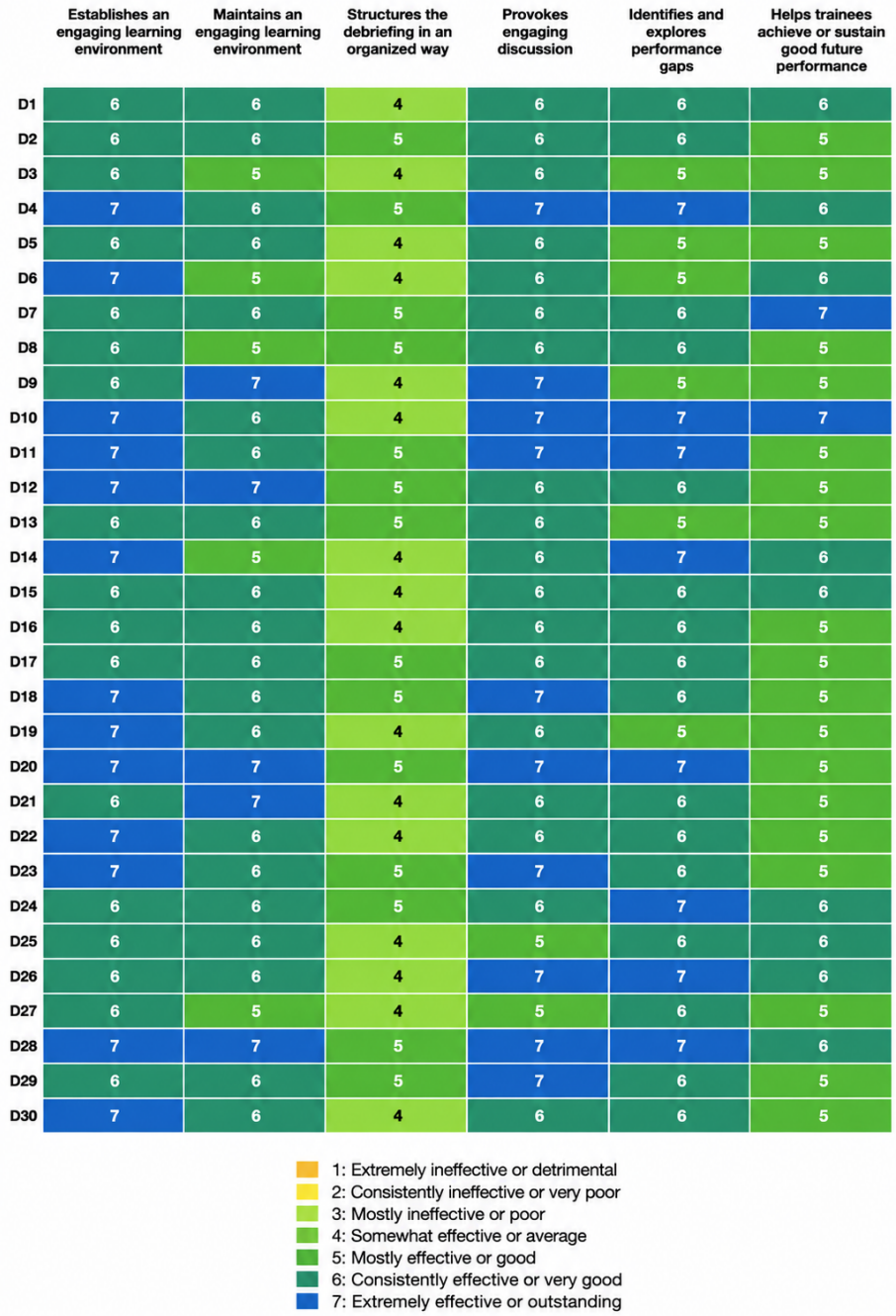
Debriefing Sample and Technical Feasibility

A total of 30 postsimulation debriefings were included in the analysis. A complete, structured AI-generated evaluation across all 6 DASH elements was obtained for all 30 (100%) debriefings, meeting the predefined feasibility outcome. All simulation sessions had complete audio recordings and usable transcripts, and no debriefings were excluded. Sessions were conducted consecutively during the study period under the supervision of a single instructor. The mean debriefing duration was 25.4 (SD 4.8; range 18-35) minutes. Transcript usability was 100%, with no technical issues or data loss observed throughout the analysis process.

DASH Ratings and Cross-Debriefing Analysis

The AI-generated analysis produced a structured qualitative evaluation and a quantitative rating using the original 7-point DASH effectiveness scale for each of the 6 DASH elements. Results are presented using a color-coded system. Figure 2 presents a heatmap of quantitative DASH ratings on the 7-point scale across the 30 debriefings and the 6 DASH elements. Detailed individual evaluations, including the systematic identification of 3 strengths and 3 areas for improvement supported by transcript excerpts, are provided in Multimedia Appendix 2.

Figure 2. Heatmap of AI-generated qualitative evaluation levels across the 30 analyzed debriefings according to the 6 dimensions of the Debriefing Assessment for Simulation in Healthcare (DASH) framework.



Each row represents an individual debriefing session, and each column represents an official DASH element. Colors correspond to the 7-point behaviorally anchored DASH rating scale, ranging from 1="extremely ineffective or detrimental" to 7="extremely effective or outstanding." Scores were generated by the LLM.

Cross-debriefing analysis revealed distinct patterns across the 6 DASH elements. "Establishes an engaging learning environment" and "Maintains an engaging learning environment" were consistently highly rated, with median scores of 6 (IQR 6-7) and 6 (IQR 6-6), respectively. "Provokes engaging discussion" and "Identifies and explores performance gaps" also received high ratings, with median scores of 6 (IQR 6-7) and 6 (IQR 6-6), respectively. In contrast, "Structures the debriefing in an organized way" received the lowest ratings, with a median score of 4 (IQR 4-5), while "Helps trainees achieve or sustain good future performance" had a median score of 5 (IQR 5-6).

Figure 3 presents a radar chart illustrating the median DASH ratings across the 6 DASH elements using the original 7-point DASH effectiveness scale. Table 1 presents the AI-generated cross-debriefing synthesis for each DASH element, including an overall evaluation and the principal strengths and areas for improvement identified across all 30 debriefings.

Scores represent the overall ratings generated by the LLM for each DASH dimension using the 7-point behaviorally anchored rating scale.

Figure 3. Radar diagram illustrating the overall quantitative profile of the 30 analyzed debriefings across the 6 dimensions of the Debriefing Assessment for Simulation in Healthcare (DASH) framework.

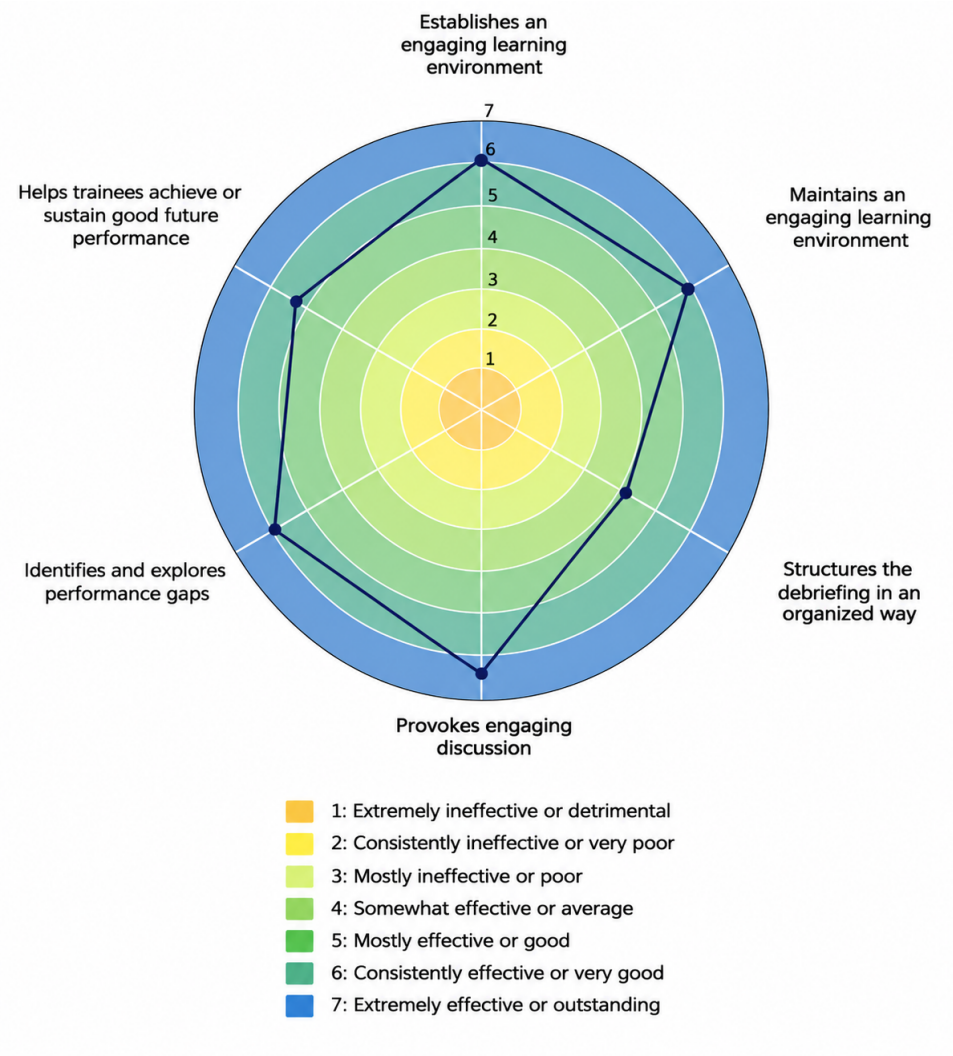


Table 1. Cross-debriefing qualitative synthesis generated by the large language model using the Debriefing Assessment for Simulation in Healthcare (DASH) framework[a].

DASH domains	DASH elements	Evaluation	Strengths	Areas for improvement
Domain 1	Establishes an engaging learning environment	Consistently effective or very good	• Psychologically safe environment that encouraged expression (eg, "I'm not sure...") and acknowledgment of errors • Mostly supportive and nonjudgmental peer feedback • Good group dynamics and active learner participation • Frequent reinforcement of appropriate behaviors (eg, calling for help and crisis management)	• Tendency toward an overly consensual climate, with limited pedagogical confrontation • Limited exploration of emotional aspects, with stress and feelings rarely explored • Uneven participation among group members • Climate often focused on cognitive analysis rather than lived experience
Domain 2	Maintains an engaging learning environment	Consistently effective or very good	• Focus generally maintained on core clinical issues	• Frequent digressions, including anecdotes and peripheral or off-topic discussions

DASH domains	DASH elements	Evaluation	Strengths	Areas for improvement
			• Implicit use of structuring frameworks (eg, the airway, breathing, and circulation approach and clinical reasoning) • Ability to return to the case after digressions • Relevant discussions on real-life management	• Occasional dilution of the educational framework during prolonged exchanges • Transitions between debriefing phases often not explicit • Difficulty maintaining a consistent thread
Domain 3	Structures the debriefing in an organized way	Somewhat effective or average	• Implicit organization around the clinical case and its progression • Logical progression in analysis (events → reasoning → actions) • Identification of key moments in management	• Near absence of formal structure, with no explicit phases • Occasionally nonlinear flow with frequent backtracking • Lack of final synthesis in most debriefings • High variability across sessions
Domain 4	Provokes engaging discussion	Consistently effective or very good	• Strong interaction among participants in most debriefings • Active coconstruction of clinical reasoning • Well-argued debates, particularly regarding critical decisions • Ability to generate multiple hypotheses	• Limited facilitation or structuring of discussions by the instructor • Discussions sometimes lengthy and highly technical • Risk of cognitive overload • Some interesting ideas not fully explored
Domain 5	Identifies and explores performance gaps	Consistently effective or very good	• Frequent and relevant identification of errors and difficulties • Analysis of clinical reasoning (including cognitive biases and decision-making) • Identification of critical moments in patient management • Ability to discuss alternative approaches	• Prioritization of performance gaps rarely made explicit • Some gaps remain implicit or are only partially analyzed • Links between actions and consequences sometimes insufficiently developed • Variability in the depth of analysis across sessions
Domain 6	Helps trainees achieve or sustain good future performance	Mostly effective or good	• Frequent and relevant educational input • Introduction of transferable strategies (eg, crisis resource management, anticipation, and communication) • Rich sharing of clinical expertise • Key messages sometimes highly actionable in critical situations	• Few concrete and individualized action plans • Recommendations often general • Frequent absence of a practice-oriented final synthesis • High variability across debriefings

[a]The table summarizes the dominant strengths, recurring areas for improvement, and overall qualitative ratings identified across the 30 analyzed debriefings.

Discussion

Principal Findings

In this observational study, we demonstrated that an LLM (ChatGPT) can generate a structured quantitative and qualitative analysis of postsimulation debriefings using the original DASH framework. The analysis was successfully applied to all 30 debriefings without any exclusion or technical issues, supporting the operational feasibility of this approach in a real-world simulation training setting.

The findings highlight the model's ability to generate structured evaluations across the 6 DASH elements. The resulting profiles differentiated areas of consistently high performance from elements with greater potential for improvement. Beyond quantitative ratings, the qualitative framework used in this study enabled the systematic identification of strengths and areas for improvement for each element, providing a more nuanced characterization of debriefing practices and educational interactions.

The cross-debriefing analysis revealed recurring patterns in debriefing practices across sessions. Four DASH elements—"Establishes an engaging learning environment," "Maintains an engaging learning environment," "Provokes engaging discussion," and "Identifies and explores performance gaps"—received median ratings of 6 (IQR 6-7, 6-6, 6-7,

and 6-6, respectively), reflecting generally effective debriefing practices across these areas. However, particularly for elements related to the learning environment, consistently high ratings may also reflect the LLM's limited ability to discriminate between performance levels, aspects partly conveyed through nonverbal or paraverbal cues that were not captured in the written transcripts. In contrast, "Structures the debriefing in an organized way" received the lowest median rating (4, IQR 4-5), while "Helps trainees achieve or sustain good future performance" received a median rating of 5 (IQR 5-6). These findings suggest that debriefing structure and the translation of reflection into strategies for future performance may represent areas with greater potential for improvement. These findings are consistent with the previous literature highlighting the inherent complexity of standardizing debriefing practices despite the availability of structured frameworks. Kolbe et al [22,23] identified several challenges associated with team debriefing, including information-sharing biases and threats to psychological safety. Previous work has also demonstrated that specific communication patterns, such as instructor questioning strategies and participant contributions, are positively associated with learning outcomes [24]. Grant et al [25] further described challenging situations that may influence the quality of the learning environment during debriefing.

Comparison With Previous Work

These findings align with the growing body of literature exploring the integration of AI into medical education, particularly for the analysis of educational interactions and the generation of structured automated feedback [26]. The use of a validated framework such as DASH provides a strong methodological foundation by aligning model-generated outputs with recognized educational standards. Within this perspective, AI may not simply automate assessment processes but may also contribute to reflective practice, faculty development, and continuous improvement in simulation-based education. Rather than replacing human expertise, such approaches may serve as complementary tools capable of supporting instructor reflection and enhancing consistency in debriefing evaluation [23,27]. Similarly, Vaughn et al [28] described a generative AI feedback interface designed to provide structured feedback to simulation facilitators on their debriefing practices, further supporting the potential role of AI as an adjunct to faculty development.

In a study involving second-year medical students completing a written pharmacotherapy assignment on anticoagulant prescribing, Ali et al [29] compared unedited ChatGPT-generated feedback with feedback provided by human tutors. Human-generated feedback was rated significantly higher in terms of clarity, relevance, actionability, comprehensiveness, accuracy, and overall usefulness, although 62.3% of students considered the 2 forms of feedback complementary [29]. Similarly, a randomized controlled trial involving 129 first-year medical students compared ChatGPT-generated feedback with expert-written feedback on clinical reasoning questions of increasing complexity. Differences were particularly apparent for more complex clinical scenarios, such as complicated urinary tract infections, suggesting limitations in the ability of AI-generated feedback to capture clinical nuance and supporting the continued role of expert review [30]. Furthermore, Boscardin et al [31] identified algorithmic bias and transparency issues as major challenges requiring ethical guidelines for the implementation of AI in medical education. Finally, the risk of excessive cognitive reliance on generative AI has been raised, potentially undermining learners' cognitive autonomy without appropriate curricular integration emphasizing ethical use and critical appraisal [32].

Limitations

This study should nevertheless be interpreted within the context of its observational design. The analysis relies exclusively on textual transcripts and therefore did not capture nonverbal dimensions of debriefing such as tone, pauses, body language, or emotional dynamics, all of which may contribute meaningfully to pedagogical interventions and learner support [33,34]. Although the model was instructed to base its evaluations and supporting excerpts exclusively on the provided transcripts, no systematic excerpt-by-excerpt verification of AI-generated outputs against the source transcripts or complete audio recordings was performed. Therefore, occasional inaccuracies or speaker misattributions cannot be entirely excluded. Furthermore, no formal assessment of potential AI bias, including bias related to gender, race, or communication style, was performed. Given the known susceptibility of LLMs to hallucination and bias, these issues should be specifically addressed in future validation studies. In addition, all debriefings were conducted by a single instructor, limiting the transferability of the findings to other educational contexts and facilitation styles. This design does not establish whether the model can consistently assess debriefings conducted by instructors with different facilitation styles or levels of performance. Moreover, the use of a single instructor did not eliminate session-level heterogeneity. Different simulation scenarios were used, groups could include residents from different postgraduate years, and variations in participant performance, team interactions, and scenario progression may have influenced the content and structure of individual debriefings. Prior studies have shown that structured debriefing tools can substantially influence facilitator discourse and debriefing approaches, underscoring the importance of instructor-related variability in simulation-based education [35]. Finally, this study did not include a comparison with human DASH-trained raters; therefore, the validity, accuracy, and reliability of model-generated assessments relative to expert human evaluation cannot be established from the present findings.

Future Directions and Conclusions

The next step in this work will be to compare AI-generated evaluations with those of a panel of simulation experts trained in the use of the DASH framework. Such a comparison will allow us to assess the level of agreement between human and automated analysis, identify the most robust or automation-sensitive elements, and clarify the potential role of these tools in the pedagogical evaluation of debriefings.

In conclusion, this observational study suggests that an LLM can generate structured quantitative and qualitative analyses of postsimulation debriefings using the original DASH framework in a real-world training context. These findings provide a basis for the development of tools to support evaluation and faculty development in simulation-based education.

Acknowledgments

JP acknowledge the French Society of Anesthesiology and Intensive Care for its support through a mobility grant awarded to fund a simulation fellowship at the University of Montreal.

Funding

The authors declare that this research received no specific funding from public, commercial, or nonprofit funding agencies.

Data Availability

The datasets generated and analyzed during this study are not publicly available due to confidentiality considerations related to the simulation debriefing transcripts. Deidentified data may be made available from the corresponding author upon reasonable request, subject to institutional and ethical approval.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Complete standardized prompt used for AI-based analysis of postsimulation debriefings.
[DOCX File (Microsoft Word File), 18 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Detailed qualitative evaluations of the 30 debriefings across the 6 Debriefing Assessment for Simulation in Healthcare (DASH) dimensions, including strengths, areas for improvement, and illustrative transcript excerpts.
[DOCX File (Microsoft Word File), 4367 KB-Multimedia Appendix 2]

References

1. Elendu C, Amaechi DC, Okatta AU, et al. The impact of simulation-based training in medical education: a review. Medicine (Baltimore). Jul 5, 2024;103(27):e38813. [doi: 10.1097/MD.0000000000038813] [Medline: 38968472]
2. Dreifuerst KT. Using debriefing for meaningful learning to foster development of clinical reasoning in simulation. J Nurs Educ. Jun 2012;51(6):326-333. [doi: 10.3928/01484834-20120409-02] [Medline: 22495923]
3. Levett-Jones T, Lapkin S. A systematic review of the effectiveness of simulation debriefing in health professional education. Nurse Educ Today. Jun 2014;34(6):e58-e63. [doi: 10.1016/j.nedt.2013.09.020] [Medline: 24169444]
4. Lee-Jayaram J, Ishikawa K, Lee YJ, et al. Localizing and tailoring the debriefing assessment for simulation in healthcare to optimize fit. Cureus. Dec 2024;16(12):e75570. [doi: 10.7759/cureus.75570] [Medline: 39803164]
5. Arora S, Ahmed M, Paige J, et al. Objective structured assessment of debriefing: bringing science to the art of debriefing in surgery. Ann Surg. Dec 2012;256(6):982-988. [doi: 10.1097/SLA.0b013e3182610c91] [Medline: 22895396]
6. Hull L, Russ S, Ahmed M, Sevdalis N, Birnbach DJ. Quality of interdisciplinary postsimulation debriefing: 360° evaluation. BMJ Simul Technol Enhanc Learn. 2017;3(1):9-16. [doi: 10.1136/bmjstel-2016-000125] [Medline: 35515097]
7. Bae J, Lee J, Jang Y, Lee Y. Development of simulation education debriefing protocol with faculty guide for enhancement clinical reasoning. BMC Med Educ. Jun 11, 2019;19(1):197. [doi: 10.1186/s12909-019-1633-8] [Medline: 31185972]
8. Brett-Fleegler M, Rudolph J, Eppich W, et al. Debriefing assessment for simulation in healthcare: development and psychometric properties. Simul Healthc. Oct 2012;7(5):288-294. [doi: 10.1097/SIH.0b013e3182620228] [Medline: 22902606]
9. Zargham S, Hanson A, Laniewicz M, et al. Psychometric testing of the Debriefing Assessment for Simulation in Healthcare (DASH) for trainee-led, in situ simulations in the pediatric emergency department context. AEM Educ Train. 2020;5(2):e10482. [doi: 10.1002/aet2.10482] [Medline: 33842804]
10. Cheng A, Grant V, Dieckmann P, Arora S, Robinson T, Eppich W. Faculty development for simulation programs: five issues for the future of debriefing training. Simul Healthc. Aug 2015;10(4):217-222. [doi: 10.1097/SIH.0000000000000090] [Medline: 26098492]
11. Chan MM, Wan AW, Cheung DS, et al. Integration of artificial intelligence in nursing simulation education: a scoping review. Nurse Educ. 2025;50(4):195-200. [doi: 10.1097/NNE.0000000000001851] [Medline: 40168182]

12. Gordon M, Daniel M, Ajiboye A, et al. A scoping review of artificial intelligence in medical education: BEME guide no. 84. Med Teach. Apr 2024;46(4):446-470. [doi: 10.1080/0142159X.2024.2314198] [Medline: 38423127]
13. Tozsin A, Ucmak H, Soyturk S, et al. The role of artificial intelligence in medical education: a systematic review. Surg Innov. Aug 2024;31(4):415-423. [doi: 10.1177/15533506241248239] [Medline: 38632898]
14. Rincón EH, Jimenez D, Aguilar LA, Flórez JM, Tapia ÁE, Peñuela CL. Mapping the use of artificial intelligence in medical education: a scoping review. BMC Med Educ. Apr 12, 2025;25(1):526. [doi: 10.1186/s12909-025-07089-8] [Medline: 40221725]
15. Gonzalez L, Nagendran A. Artificial intelligence (AI)-facilitated debriefing: a pilot study. Clin Simul Nurs. Aug 2025;105:101782. [doi: 10.1016/j.ecns.2025.101782]
16. Henquet M, Bellucci T. AI-assisted debrief: automated flight debriefing summarization and competency assessment. In: Proceedings of the 2nd International Conference on Cognitive Aircraft Systems - Volume 1: ICCAS. SciTePress; 2024:72-79. [doi: 10.5220/0012970200004562]
17. Echeverria V, Zhao L, Alfredo R, et al. TeamVision: an AI-powered learning analytics system for supporting reflection in team-based healthcare simulation. In: CHI '25: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery; 2025:1-22. [doi: 10.1145/3706598.3713395]
18. Fung TC, Chan SL, Lam CF, et al. Effects of generative artificial intelligence (GenAI) patient simulation on perceived clinical competency among global nursing undergraduates: a cross-over randomised controlled trial. BMC Nurs. Jul 17, 2025;24(1):934. [doi: 10.1186/s12912-025-03492-0] [Medline: 40676632]
19. Hong E, Kazmir S, Dylik B, et al. Exploring the use of a large language model in simulation debriefing: an observational simulation-based pilot study. Simul Healthc. 2025;20(6):366-371. [doi: 10.1097/SIH.0000000000000861] [Medline: 40326794]
20. Hicke Y, Geathers J, Vu K, et al. MedSimAI: simulation and formative feedback generation to enhance deliberate practice in medical education. In: LAK '26: Proceedings of the LAK26: 16th International Learning Analytics and Knowledge Conference. Association for Computing Machinery; 2026:415-424. [doi: 10.1145/3785022.3785092]
21. Sumpter S. Automated generation of high-quality medical simulation scenarios through integration of semi-structured data and large language models. arXiv. Preprint posted online on Apr 30, 2024. [doi: 10.48550/arXiv.2404.19713]
22. Kolbe M, Grande B, Spahn DR. Briefing and debriefing during simulation-based training and beyond: content, structure, attitude and setting. Best Pract Res Clin Anaesthesiol. Mar 2015;29(1):87-96. [doi: 10.1016/j.bpa.2015.01.002] [Medline: 25902470]
23. Kolbe M, Grande B, Lehmann-Willenbrock N, Seelandt JC. Helping healthcare teams to debrief effectively: associations of debriefers' actions and participants' reflections during team debriefings. BMJ Qual Saf. Mar 2023;32(3):160-172. [doi: 10.1136/bmjqs-2021-014393] [Medline: 35902231]
24. 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. Med Teach. Dec 2021;43(12):1381-1390. [doi: 10.1080/0142159X.2021.1948521] [Medline: 34260335]
25. Grant VJ, Robinson T, Catena H, Eppich W, Cheng A. Difficult debriefing situations: a toolbox for simulation educators. Med Teach. Jul 2018;40(7):703-712. [doi: 10.1080/0142159X.2018.1468558] [Medline: 29792100]
26. Parmar R, Lee M, Budzikowski A. Multimodal AI for automated resuscitation debriefing: feasibility of timeline reconstruction using commercially available tools. Med Teach. Jan 22, 2026:1-4. [doi: 10.1080/0142159X.2026.2618505] [Medline: 41568790]
27. Cheng A, Eppich W, Kolbe M, Meguerdichian M, Bajaj K, Grant V. A conceptual framework for the development of debriefing skills: a journey of discovery, growth, and maturity. Simul Healthc. Feb 2020;15(1):55-60. [doi: 10.1097/SIH.0000000000000398] [Medline: 31743312]
28. Vaughn J, Ford SH, Sheckells A, Fasnacht H, Crawford D, Dogan G. Enhancing simulation facilitator debriefing using a generative artificial intelligence feedback interface and retrieval-augmented generation: a pilot study. Clin Simul Nurs. Dec 2025;109:101843. [doi: 10.1016/j.ecns.2025.101843]
29. Ali M, Harbieh I, Haider KH. Bytes versus brains: a comparative study of AI-generated feedback and human tutor feedback in medical education. Med Teach. Jan 2026;48(1):131-141. [doi: 10.1080/0142159X.2025.2519639] [Medline: 40528415]
30. Çiçek FE, Ülker M, Özer M, Kıyak YS. ChatGPT versus expert feedback on clinical reasoning questions and their effect on learning: a randomized controlled trial. Postgrad Med J. Apr 22, 2025;101(1195):458-463. [doi: 10.1093/postmj/qgae170] [Medline: 39656920]
31. Boscardin CK, Abdulnour RE, Gin BC. Macy Foundation innovation report part I: current landscape of artificial intelligence in medical education. Acad Med. Sep 1, 2025;100(9S Suppl 1):S15-S21. [doi: 10.1097/ACM.0000000000006107] [Medline: 40456178]

32. Izquierdo-Condoy JS, Arias-Intriago M, Tello-De-la-Torre A, Busch F, Ortiz-Prado E. Generative artificial intelligence in medical education: enhancing critical thinking or undermining cognitive autonomy? J Med Internet Res. Nov 3, 2025;27:e76340. [doi: 10.2196/76340] [Medline: 41183320]
33. Ross S. Twelve tips for effective simulation debriefing: a research-based approach. Med Teach. Jun 2021;43(6):642-645. [doi: 10.1080/0142159X.2020.1831689] [Medline: 33053315]
34. Coggins A, Zaklama R, Szabo RA, et al. Twelve tips for facilitating and implementing clinical debriefing programmes. Med Teach. May 2021;43(5):509-517. [doi: 10.1080/0142159X.2020.1817349] [Medline: 33032476]
35. Thompson R, Sullivan S, Campbell K, Osman I, Statz B, Jung HS. Does a written tool to guide structured debriefing improve discourse? Implications for interprofessional team simulation. J Surg Educ. Nov 2018;75(6):e240-e245. [doi: 10.1016/j.jsurg.2018.07.001] [Medline: 30093336]

Abbreviations

AIDE: Artificial Intelligence for Debriefing Evaluation
CEREP: Comité d'éthique de la recherche en éducation et en psychologie
DASH: Debriefing Assessment for Simulation in Healthcare
LLM: large language model

Edited by Amy Bucher; peer-reviewed by Isabella Menchetti, Jannet J Lee-Jayaram; submitted 19.May.2026; final revised version received 14.Aug.2026; accepted 20.Sep.2026; published 17.Nov.2026

Please cite as:
Pico J, Bouhadiba T, Jazestani S, Zhu E, Nguyen BK, Baroudi R, Dababneh S, Tanoubi I
Use of Large Language Models for Structured Quantitative and Qualitative Analysis of Simulation Debriefings Using the DASH Framework (AIDE Study): Prospective Observational Study
JMIR Form Res 2026;10:e101797
URL: https://formative.jmir.org/2026/1/e101797
doi: 10.2196/101797

© Julien Pico, Tarikeddine Bouhadiba, Shiva Jazestani, Eric Zhu, Ba-Khang Nguyen, Reda Baroudi, Said Dababneh, Issam Tanoubi. Originally published in JMIR Formative Research (https://formative.jmir.org), 17.Nov.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.