education sciences

MDPI

Article

Bot or Not? Differences in Cognitive Load Between Human- and Chatbot-Led Post-Simulation Debriefings

Dominik Evangelou *, Miriam Mulders and Kristian Heinrich Träg

Chair of Educational Technology and Instructional Design, University of Duisburg-Essen, Universitätsstraße 2, 45141 Essen, Germany; miriam.mulders@uni-due.de (M.M.); kristian.traeg@uni-due.de (K.H.T.)
* Correspondence: dominik.evangelou@uni-due.de

Abstract

Understanding how different debriefing formats impact learner's cognitive load is crucial for designing effective post-simulation reflection activities. This paper examines cognitive load after post-simulation debriefings facilitated either by a human instructor or a generative AI Chatbot. In a controlled study with $N = 45$ educational science students, 23 participants engaged in a lecturer-facilitated debriefing, while 22 completed a chatbot-guided session. Cognitive load was assessed across intrinsic, extraneous, and germane dimensions. Results revealed no statistically significant differences between the two debriefing methods. Future research should examine AI-led debriefings with larger samples and employ complementary measures of cognitive load to provide a more comprehensive understanding.

Keywords: debriefing; AI; generative chatbot; virtual reality; cognitive load

1. Introduction

Research on simulation-based learning suggests that structured debriefing sessions support effective learning with simulations by facilitating guided reflection on learners' actions and experience (Shinnick et al., 2011; Ryoo & Ha, 2015; Crookall, 2023). Even the most sophisticated and immersive simulations reach their full educational potential only when participants are provided with guided opportunities to critically analyze their decisions and actions (Crookall, 2023). Debriefing serves as this essential phase by helping learners interpret their experiences, consider alternative strategies, and connect their learning to professional practice (Kriz & Nöbauer, 2015; Luctkar-Flude et al., 2021). Despite its importance, debriefing—particularly within Virtual Reality (VR) and similar simulation environments—remains a resource-intensive practice, often limited by the availability of qualified facilitators (INACSL Standards Committee, 2016; Metcalfe et al., 2007). Research on simulation debriefing is expanding but remains largely fragmented across domains, with a predominant focus on healthcare contexts (Cheng et al., 2020; Favolise, 2024; Garden et al., 2015). As such, there is an urgent need to extend empirical and conceptual work into broader educational settings. Two primary debriefing strategies are prominent in the literature: facilitator-led and self-guided approaches. These differ in instructional design, group size, technological support, and instructor (Dufrene & Young, 2014; Luctkar-Flude et al., 2021). Although facilitator-led debriefing is endorsed by professional bodies such as the International Nursing Association for Clinical Simulation and Learning (INACSL) (INACSL Standards Committee, 2016), empirical evidence offering unequivocal support for its superiority remains limited (Dufrene & Young, 2014). This gap highlights the necessity

Check for updates

Academic Editors: Ali Darejeh and Sara Mashayekh

Received: 9 December 2025
Revised: 30 January 2026
Accepted: 3 February 2026
Published: 6 February 2026

https://doi.org/10.3390/educsci16020255

for systematic comparison across debriefing modalities. Recent advances in generative artificial intelligence (AI), particularly with large language model (LLM)-based chatbots, present promising opportunities to augment debriefing at scale. These AI-driven conversational agents may simulate interactive debriefing dialogs analogous to human facilitators potentially alleviating resource constraints in simulation-based education (Evangelou et al., 2025, 2026).

Alongside its role in supporting effective knowledge acquisition, debriefing also raises the question of whether and to what extent post-simulation reflection may affect learner's cognitive load. Cognitive Load Theory (CLT) posits that the efficiency of learning largely depends on the balance between the cognitive demands imposed by a task and the limited capacity of working memory (Sweller, 1988). When instructional designs minimize extraneous load and foster germane processing, learners are more likely to construct and retain meaningful knowledge structures (Sweller et al., 2019).

The aim of this study is to compare cognitive load resulting from AI-led versus moderator-led post-simulation debriefings among students in a VR simulation context to determine whether chatbot facilitation affects intrinsic, extraneous, and germane cognitive load differently.

2. Theoretical Background

2.1. Debriefing

As highlighted in the introduction, debriefing is a fundamental component of simulation-based learning, essential for supporting reflection that deepens and consolidates learners' knowledge (Dreifuerst, 2015; Fey & Jenkins, 2015). It creates space for learners to critically review their actions, consider alternative approaches, and relate their simulated experience to theoretical concepts and professional practice (Sawyer et al., 2016). According to Kolb's (2014) experiential learning theory, this reflective process is key to transforming concrete experience into abstract understanding. Effective debriefing encourages learners to identify emotional reactions, assess their decisions, and derive transferable insights (Rudolph et al., 2006), thereby fostering both cognition growth and professional confidence (Cantrell, 2008).

Qualitative studies reveal that a complex situation alone is insufficient for competence development; instructional conditions such as well-structured debriefing sessions are essential for transferring learning to practice (Hense & Kriz, 2008; Kriz et al., 2007). Empirical research on debriefing often focuses on comparing methods or exploring learner's experiences (Dufrene & Young, 2014). For example, learners who participate in post-simulation debriefing demonstrate greater knowledge gains than those who do not (Shinnick et al., 2011). Debriefing may be structured and supported through various approaches, including instructor guidance, peer feedback (Boet et al., 2011), and the use of supporting tools such as video-assisted sessions to facilitate reflection (Chronister & Brown, 2012; Grant et al., 2010). Together, these studies underline structured debriefing's critical role in enhancing knowledge acquisition and professional competence.

Debriefing formats range from facilitator-led to self-guided approaches. Facilitated debriefings involve trained moderators guiding reflection, providing feedback, and helping learners interpret key simulation events (Cheng et al., 2017; Sawyer et al., 2016). This approach is widely seen as best practice because it tailors discussions to learner needs, resolves misunderstandings, and models reflection skills (INACSL Standards Committee, 2016; Fey & Jenkins, 2015). Established frameworks like Debriefing with Good Judgment (Rudolph et al., 2006) and PEARLS (Eppich & Cheng, 2015) emphasize psychological safety, respect, and learner-centered dialog as foundations for effective debriefing.

However, moderated debriefing demands significant time and skilled facilitators, limiting scalability especially in large educational context (Cheng et al., 2017). Self-guided formats, which provide prompts or structured guides for independent reflection, have emerged as scalable alternatives (Boet et al., 2014; Tosterud et al., 2013). While promoting autonomy, their effectiveness is debated. Novices, in particular, may struggle with unguided reflection due to limited metacognitive skills (Dufrene & Young, 2014). Still, with adequate scaffolding through prompts or digital support, self-guided debriefing can complement or substitute facilitated sessions (Evangelou et al., 2026; Koole et al., 2012; Luctkar-Flude et al., 2021).

2.2. Cognitive Load

Cognitive Load Theory (CLT) explains how the limited capacity of working memory constrains learning processes and learning outcomes. Learning is most effective when cognitive resources are not overloaded, allowing learners to allocate cognitive effort to processes that foster understanding and knowledge construction (Sweller, 1988; Sweller et al., 2019).

Traditionally, CLT distinguishes between three types of cognitive load (Chandler & Sweller, 1991; Paas et al., 2003). Intrinsic cognitive load refers to the inherent complexity of the learning material, which is determined by task characteristics and learners' prior knowledge. Extraneous cognitive load arises from suboptimal instructional design, such as unclear instructions or irrelevant information, and consumes working-memory resources without contributing to learning. Germane cognitive load reflects the cognitive effort learners invest in schema construction and meaningful processing.

In later theoretical developments, the conceptual status of germane cognitive load has been critically discussed. In particular, it has been argued that germane load does not constitute a separate type of cognitive load, but rather reflects the effective use of working-memory resources devoted to intrinsic load (Sweller, 2010, 2011; Kalyuga, 2011). Despite this ongoing debate, the three-component framework continues to be widely used in empirical research, especially in studies employing established self-report instruments that operationalize intrinsic, extraneous, and germane cognitive load as distinct dimensions (Paas et al., 2003; Leppink et al., 2013; Klepsch et al., 2017). Accordingly, the present study adopts this operationalization to enable a differentiated examination of learners' cognitive load during post-simulation debriefing.

From an instructional perspective, effective learning environments should aim to manage intrinsic demands, minimize extraneous load, and support cognitive processes related to schema construction (Chandler & Sweller, 1991). In the context of this study, the focus lies on post-simulation debriefings as structured reflection activities. Specifically, this study examines how different facilitation formats influence learners' cognitive load in order to derive implications for the design of instructional debriefing practices.

2.3. Debriefing and Cognitive Load

Research investigating the relationship between debriefing and learners' perceived cognitive load remains limited. Existing studies predominantly utilize a pre–post design, assessing cognitive load immediately after the VR experience and again following debriefing (Fraser & McLaughlin, 2019; Miller et al., 2025). These studies consistently report that total cognitive load measured post-scenario is moderately high and increases significantly after debriefing. Moreover, elevated cognitive load following debriefing has been linked to reduced tranquility (Fraser & McLaughlin, 2019). The choice of cognitive load measurement instrument appears to critically influence findings. Miller et al. (2025), compared the Paas scale (Paas et al., 2003) and the CLAS-Sim instrument (Greer et al., 2023) in a study where participants completed ten VR scenarios, each followed by a 15 min debriefing. Their results indicate that germane cognitive load reported via CLAS-Sim (Greer et al., 2023) after debriefing was higher compared to the Paas scale (Paas et al., 2003). They further argue that germane cognitive load should primarily be measured post-debriefing to capture its full manifestation. In contrast, intrinsic and extraneous cognitive load measurements showed no differences between instruments, both pre and post debriefing, likely reflecting the stable task complexity inherent in the simulation. A study comparing cognitive load across debriefing methods indicates that there are no significant differences between peer-led and instructor-led debriefings (Na & Roh, 2021). Furthermore, the authors also reported that total cognitive load was higher after debriefing compared to before. Similarly, a separate study examining video-assisted versus non-video-assisted debriefings found no significant differences in cognitive load between approaches either (Braund et al., 2025).

Chatbots as Debriefers

Conversational agents have recently been used in educational contexts to support learning, feedback and reflection (Zawacki-Richter et al., 2019; Kasneci et al., 2023). Advances in natural language processing and generative AI enable adaptive, context-aware dialogs resembling human tutoring (Winkler & Soellner, 2018). Their applications include language learning, tutoring, assessment, and emotional support (Holmes et al., 2019; Kerlyl et al., 2007). In simulation-based learning, chatbots offer scalable alternatives to human facilitators, especially for post-simulation debriefings (Kumar et al., 2025; Zhu et al., 2025). They guide structured reflection, pose questions, and provide empathetic feedback (Ortega-Ochoa et al., 2024). Evangelou et al. (2025) demonstrated that generative chatbots can effectively maintain their role as debriefers even in complex conversations, with learners using the chatbot to reflect and occasionally take dialog control.

Recent generative AI models produce coherent, contextually relevant responses, supporting metacognitive tasks like reflection and self-assessment (Kasneci et al., 2023; Zawacki-Richter et al., 2019). Early evidence shows that learners find AI-based debriefings engaging and helpful when pedagogically designed with learner-centered dialog (Nghi & Anh, 2024; Wang & Akhter, 2025). Limitations remain, including challenges recognizing emotional cues and handling ambiguous responses (Liang & Hwang, 2023).

2.4. Research Questions and Hypotheses

The increasing importance of AI in education (Ifenthaler et al., 2024), combined with challenges such as limited human resources in higher education (McDonald, 2013), highlights the potential of AI-facilitated reflective conversations, including debriefings. However, there are concerns that AI-driven facilitation may impose additional cognitive load that could overwhelm learners (Memarian & Doleck, 2023; Klar, 2025). Importantly, the majority of existing research on debriefing is concentrated within the healthcare domain, with relatively few studies directly comparing cognitive load across different debriefing methods (Na & Roh, 2021). This contextual gap underscores the significance of the present study. Accordingly, we investigate whether AI-led post-simulation debriefings produce cognitive load profiles comparable to those of moderator-led debriefings recommended by the INACSL Standards Committee (2016). From this, the following hypotheses are derived:

1. Perceived intrinsic cognitive load will be higher after a chatbot-led debriefing compared to a moderator-led debriefing.
2. Perceived extraneous cognitive load will be higher after a chatbot-led debriefing compared to a moderator-led debriefing.
3. Perceived germane cognitive load will be lower after a chatbot-led debriefing compared to a moderator-led debriefing.

3. Methods

The study is part of the research project *VR-Hybrid* which investigates the potential of VR for training counseling techniques in higher education. The focus of the present analysis is on learners' intrinsic, extraneous, and germane cognitive load following a post-simulation debriefing session. The VR training took place in January 2025 in a laboratory setting, using a Meta Quest 3 and the VR platform Engage. The training was embedded within a university seminar in educational sciences, in which students were introduced to counseling techniques and strategies.

3.1. Participants

The study involved undergraduate students enrolled in education science programs at a large German university. Initially, *N* = 46 students participated, but one was excluded after withdrawing due to motion sickness during the intervention, resulting in a final sample of *N* = 45. The majority of the participants were women (*n* = 39, 86.7%), with an average age of 24.5 years (*SD* = 8.7). Recruitment took place through announcements in relevant courses, and participation was voluntary in exchange for course credit. Written informed consent was obtained from all students, and the study was approved in line with university's ethics procedures. Students were informed that they could withdraw at any time without consequence. Before starting the first questionnaire, each participant created a unique four-digit code, which enabled pseudonymized matching of responses across time points.

3.2. Procedure

Participants scheduled individual appointments for the VR training as part of one-on-one sessions. At the beginning of each session, they completed a pre-test questionnaire collecting demographic information, prior VR experience, empathy, self-efficacy, and counseling competence. Participants were then randomly assigned to one of two avatar conditions (customized vs. generic). They subsequently engaged in a VR-based counseling scenario in which they interacted with a standardized client ("Lena") in a virtual counseling room. The role of the client was enacted by trained research assistants, who were located in separate rooms and did not meet the participants in person.

Following the VR training, participants completed a second questionnaire measuring self-efficacy, counseling competence, presence, realism, embodiment, perspective-taking, cognitive load related to the VR training, and engagement. Participants then took part in a debriefing session that was either facilitated by a human moderator or conducted using a generative chatbot. Both debriefing formats followed the Structured Debriefing in Simulation-Based Education framework (Palaganas et al., 2016) and employed an identical three-phase structure consisting of a reaction phase, an understanding phase, and a summary phase.

In the chatbot-guided debriefing condition, participants engaged in an individual reflection process supported by an AI-based conversational agent. Prior to the interaction, participants were provided with a printed instruction sheet outlining the purpose of the debriefing and five core reflection prompts designed to structure the reflection process (e.g., focusing on successful strategies, challenges encountered, and lessons learned). The instruction sheet remained available throughout the debriefing. Participants then interacted individually with the chatbot via a laptop placed next to the instruction sheet.

During the debriefing, the chatbot assumed the role of a facilitator and guided participants through the three-phase reflection process. The debriefing began with a reaction phase addressing participants' immediate impressions and emotional responses, followed by a phase focused on analyzing and interpreting the simulated interaction, and concluded with a summary phase aimed at consolidating key insights and considering their transfer to future practice. Throughout the interaction, the chatbot prompted participants to elaborate on their experiences, reflect on their actions, and articulate generalizable conclusions. This procedure mirrored the structure used in the human-moderated debriefing condition (Supplementary Materials).

The instructional role and dialog structure of the chatbot were developed iteratively by the research team and refined through usability testing with trained student assistants prior to the study. The chatbot was implemented using the Meta Llama 3.1 8B Instruct language model, selected for its stable and consistent conversational behavior in instructional contexts. To balance response coherence and generative flexibility, both temperature and nucleus sampling parameters were set to 0.5. A standardized system prompt defined the chatbot's role as a debriefing facilitator and ensured a consistent progression through the reflection phases across participants. A qualitative analysis of the anonymized chat transcripts as well as a quantitative examination of learners' perceptions of the debriefing process and its effectiveness are reported in related publications (Evangelou et al., 2025, 2026).

Finally, participants completed a post-debriefing questionnaire assessing self-efficacy and counseling competence, as well as perceptions of the debriefing and cognitive load specific to the debriefing phase.

3.3. Measurement Instruments

For the purposes of this paper, only the measurement instrument assessing cognitive load is relevant. Cognitive load was measured using the questionnaire by Klepsch et al. (2017), assessing intrinsic (ICL), extraneous (ECL), and germane cognitive load (GCL). The instrument comprises two items for ICL (e.g., *For this task, many things needed to be kept in mind simultaneously*), three items for GCL (e.g., *I made an effort, not only to understand several details, but to understand the overall context*), and three items for ECL (e.g., *The design of this task was very inconvenient for learning*). Responses were given on a 7-point Likert scale ranging from 1 (*strongly disagree*) to 7 (*strongly agree*). For the present study, all items were slightly adapted to explicitly refer to the debriefing phase. This adaptation was made to ensure that participants clearly associated the items with the debriefing activity rather than with the preceding simulation or the overall learning task, thereby reducing potential ambiguity in item interpretation. For example, one of the ICL items was reworded as *For the debriefing, many things needed to be kept in mind simultaneously*.

4. Results

This section reports on the quantitative results of our study. The prerequisites and descriptive statistics are described first, followed by the representation of the results in relation to the hypotheses. All analyses were conducted in R version 4.4.2. To evaluate the hypotheses, multiple *t*-tests were performed.

4.1. Prerequisites

Prior to hypothesis testing, we assessed normality using the Shapiro–Wilk test (Table 1) and homogeneity of variances using Levene's test (Table 2). For ECL, the Shapiro–Wilk test indicated a significant deviation from normality ($W = 0.89$, $p < 0.001$). In contrast, the ICL and GCL did not deviate significantly from a normal distribution. Levene's tests showed no evidence of unequal variances between conditions for any of the three measures. Consequently, the hypotheses were tested using Welch's *t*-test for ICL and GCL, and the Mann–Whitney *U* test for ECL.

Table 1. Shapiro–Wilk test—results.

Variable	*W*	*p*
ICL	0.96	0.091
ECL	0.89	<0.001
GCL	0.98	0.639

Table 2. Levene test—results.

Variable	*df*	*F*	*p*
ICL	1.43	0.01	0.928
ECL	1.43	0.66	0.421
GCL	1.43	0.14	0.712

To account for possible confounding effects of debriefing length, we compared session durations across conditions. The mean duration across all participants was 14.48 min ($SD = 4.20$). In the moderated debriefing (MB) condition, sessions averaged 14.6 min ($SD = 2.92$), while in the chatbot debriefing (SB) condition, the mean was 14.4 min ($SD = 5.35$). Levene's test indicated a significant difference in variance between the two groups $F(1{,}43) = 6.59$, $p = 0.014$, suggesting heterogeneity of variance. Therefore, Welch's *t*-test was applied, revealing no statistically significant difference in duration, $t(32.18) = 0.12$, $p = 0.905$.

4.2. Descriptive Statistics

Table 3 presents the descriptive statistics for all measures, along with Cronbach's Alpha values. While the Cronbach's alpha value for ECL indicates good internal consistency, the values for ICL and GCL can be described as poor (Cronbach, 1951).

Table 3. Descriptive statistics.

Variable	Min	*Md*	Max	*M*	*SD*	α
ICL	1	3	6.50	3.08	1.36	0.55
ECL	1	2	4.67	2.14	1.05	0.84
GCL	3	5	5.02	5.02	0.97	0.53

4.3. Hypotheses Testing

Considering the prerequisite analysis, Welch's *t*-tests were conducted for ICL and GCL, while a Mann–Whitney *U* test was applied for hypothesis 2 due to the violation of the normality assumption. The Welch's *t*-test revealed no statistically significant difference in ICL between the moderated and chatbot-led debriefings, $t(43) = 0.59$, $p = 0.557$. The effect size was negligible, Cohen's $d = 0.18$. A similar pattern emerged for GCL, where the *t*-test also indicated no significant difference between conditions, $t(43) = 1.40$, $p = 0.169$, with a small effect size, $d = 0.42$. The Mann–Whitney *U* test for ECL likewise showed no significant difference between debriefing methods, $U = 204$, $p = 0.267$, with a small effect size, $r = 0.17$ (see Table 4).

Table 4. Results of the *t*-tests.

Variable	Test	*df*/*U*	*p*	Effect Size
ICL	*t*	1.43	0.557	$d = 0.18$
ECL	*U*	204	0.267	$r = 0.17$
GCL	*t*	1.43	0.169	$d = 0.42$

These results are illustrated in Figure 1, which presents the distribution of ICL, ECL and GCL scores for both debriefing conditions using violin plots combined with boxplots and individual data points. The figure clearly shows the similar distribution patterns and the negligible differences in median and mean scores between the moderator-led and chatbot-led conditions. Additionally, Figure 2 provides a profile plot of the mean scores and 95% confidence intervals for the three cognitive load types across both conditions. This visualization confirms the consistent pattern of low ECL, moderate ICL, and high GCL, with overlapping confidence intervals indicating no statistically significant differences between debriefing formats.

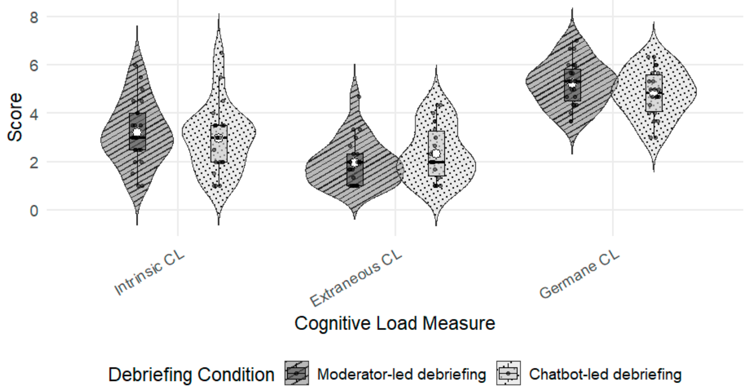

Figure 1. Cognitive load by debriefing methods.

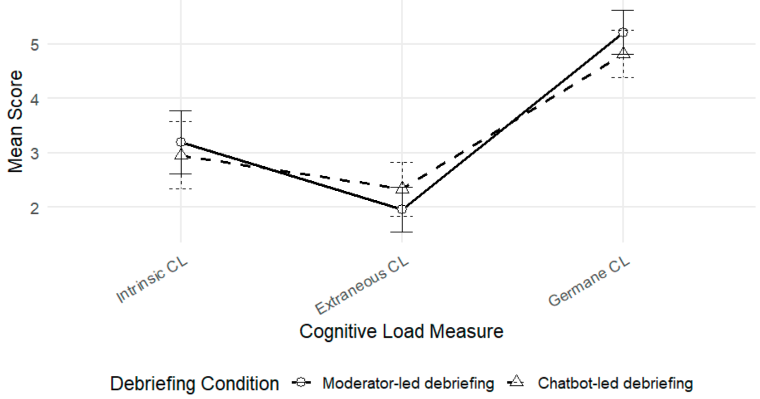

Figure 2. Profile of cognitive load across debriefing methods.

5. Discussion

The present study aimed to investigate whether chatbot-led post-simulation debriefings impose different levels of cognitive load compared to human-led debriefings. Based on CLT (Sweller, 1988; Sweller et al., 2019), three hypotheses were formulated with regard to ICL, ECL and GCL. The results, however, did not provide evidence for significant differences between the two debriefing conditions.

Hypothesis 1 predicted that ICL would be higher in the chatbot-led debriefing condition. This assumption was grounded in the idea that learners might perceive interacting with an AI system as more demanding, particularly when it comes to interpreting prompts or adapting to non-human communication patterns. Contrary to this expectation, no significant difference in ICL was observed. This finding suggests that participants did not perceive the chatbot as imposing a higher inherent complexity compared to a human moderator. A potential explanation may lie in the structured nature of the debriefing, which provided clear guidance regardless of the facilitator. Beyond these structural factors, it is theoretically conceivable that affective or social aspects of the interaction may also have played a role. For example, some learners might feel less social pressure when interacting with a non-human facilitator, whereas others may experience the absence of human responsiveness as less supportive. As these aspects were not directly measured in the present study, such interpretations remain speculative and should be addressed in future research.

Hypothesis 2 stated that ECL would be higher after chatbot-led debriefings. Such an effect was expected because technical limitations, communication breakdowns, or less adaptive responses could lead to additional processing demands. Again, the results did not support this assumption. Participants reported comparable levels of ECL in both conditions. This indicates that the chatbot was perceived as sufficiently coherent and supportive to avoid adding unnecessary distractions or confusion. From a practical perspective, this is a promising result, as it suggests that AI-driven debriefings do not inherently burden learners with avoidable processing demands.

Hypothesis 3 proposed that GCL would be lower in chatbot-led debriefings. The rationale was that human facilitators may be better able to foster reflective processing and schema construction. Yet, the findings revealed no significant differences between conditions. This suggests that the chatbot was able to support reflective engagement at a level comparable to that of a human moderator. Although the effect size indicated a small tendency towards lower GCL in the chatbot condition, the absence of significance highlights the need for further investigations with larger samples.

Taken together, the findings challenge the assumption underlying the present study that chatbot-led debriefings inherently increase learners' cognitive load. The lack of significant differences across all three dimensions of cognitive load can be interpreted positively: chatbots did not raise ICL or ECL demands in a detrimental way, nor did they significantly reduce germane processing. Thus, their use in higher education contexts can be reasonably justified, particularly in light of increasing demands for scalable teaching solutions (McDonald, 2013). These results also align with a qualitative analysis of chatbot behavior as a debriefer reported by Evangelou et al. (2025), further supporting the potential of AI-driven facilitation to effectively support reflective learning processes without imposing additional cognitive burden.

Despite these promising findings, several limitations must be acknowledged. First, the measurement of cognitive load relied on retrospective self-report scales, which, while widely used—primarily the Paas scale (Paas et al., 2003) in related studies (Braund et al., 2025; Fraser & McLaughlin, 2019; Miller et al., 2025)—have limitations in sensitivity and may not fully capture the dynamic cognitive processes during debriefing. Future research should consider alternative measurement instruments to more comprehensively assess cognitive load. Additionally, Miller et al. (2025) highlight that intrinsic and extraneous cognitive load constitute distinct constructs of GCL and emphasize the importance of evaluating these dimensions separately within simulation-based learning contexts. The relatively low reliability observed for the intrinsic and germane cognitive load subscales further constrains the interpretability of the corresponding findings, which should therefore be interpreted with caution. Second, the study was conducted with a relatively small and homogeneous sample of educational science students, which limits the generalizability of the findings to other disciplines, levels of expertise, or cultural contexts. Third, the study focused exclusively on short-term perceptions of cognitive load. Long-term effects on learning outcomes were not assessed.

Future research should therefore build on these limitations in several ways. Triangulating self-reports with physiological or behavioral measures, such as eye-tracking (Braund et al., 2025), EEG, or dual-task methods, could provide a more nuanced understanding of cognitive demands. Moreover, larger and more diverse samples are needed to examine whether chatbot-led debriefings function similarly across different learner groups and subject domains. Longitudinal studies could explore whether repeated exposure to chatbot-facilitated debriefings influences learning outcomes, reflective depth, or learner acceptance over time. Finally, qualitative approaches could shed light on learners' subjective experiences, particularly how they perceive the conversational quality and pedagogical value of chatbot interactions.

6. Conclusions

The primary aim of this study was to investigate whether chatbot-led debriefings result in higher ICL and ECL and lower GCL compared to moderator-led debriefings. Unlike studies employing a pre–post design that assess changes in cognitive load before and after debriefing (Miller et al., 2025), this study focused on directly comparing cognitive load levels between the two debriefing methods without measuring within-subject changes over time. The results provide initial evidence that chatbot-led debriefings do not significantly differ from human-led debriefings in ICL, ECL and GCL. The absence of an increase in ICL and ECL is encouraging, indicating that chatbots can be integrated into higher education settings without imposing additional cognitive burden on learners. These findings support the notion that AI-driven facilitation can serve as a practical complement to human instructors, helping to address resource constraints while maintaining instructional effectiveness.

Supplementary Materials: The following supporting information can be found at: https://www.mdpi.com/article/10.3390/educsci16020255/s1. https://osf.io/jguh4/overview?view_only=44b600c21f864e0794bb41f904f973d2, accessed on 25 November 2025, Material 1: Chatbot-led Debriefing Guide; Material 2: Prompt Chatbot.

Author Contributions: Conceptualization, D.E.; methodology, D.E., M.M., K.H.T.; formal analysis, D.E. & K.H.T.; data curation, D.E., M.M., K.H.T.; writing—original draft preparation, D.E. & M.M.; writing—review and editing, D.E., M.M. & K.H.T.; visualization, D.E.; supervision, M.M.; project administration, M.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Stiftung Innovation in der Hochschullehre, FR-124/2023.

Institutional Review Board Statement: Participation in the study was voluntary and conducted as part of a regular university course. Approval by an ethics board was not required according to the guidelines of the Interdisciplinary Center for Education Research at the University of Duisburg–Essen, as no sensitive data were collected. As stated in the institutional regulations, "There is currently no obligation to obtain an ethics vote for research projects in the field of empirical educational research (https://www.uni-due.de/izfb/ethikkommission#Aufgaben, accessed on 25 November 2025)." The experimental procedure ad-hered to the principles of the Declaration of Helsinki.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Acknowledgments: The authors acknowledge the use of artificial intelligence tools ChatGPT 5.0 for translation support.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
INACSL	International Nursing Association for Clinical Simulation and Learning
LLM	Large Language Model
CLT	Cognitive Load Theory

ICL	Intrinsic Cognitive Load
ECL	Extraneous Cognitive Load
GCL	Germane Cognitive Load

References

Boet, S., Bould, M. D., Bruppacher, H. R., Desjardins, F., Chandra, D. B., & Naik, V. N. (2011). Looking in the mirror: Self-debriefing versus instructor debriefing for simulated crises*. *Critical Care Medicine*, *39*(6), 1377–1381. [CrossRef] [PubMed]

Boet, S., Bould, M. D., Fung, L., Qosa, H., Perrier, L., Tavares, W., Reeves, S., & Tricco, A. C. (2014). Transfer of learning and patient outcome in simulated crisis resource management: A systematic review. *Canadian Journal of Anaesthesia*, *61*(6), 571. [CrossRef]

Braund, H., Hall, A. K., Caners, K., Walker, M., Dagnone, D., Sherbino, J., Sibbald, M., Wang, B., Howes, D., Day, A. G., Wu, W., & Szulewski, A. (2025). Evaluating the value of eye-tracking augmented debriefing in medical simulation—A pilot randomized controlled trial. *Simulation in Healthcare: The Journal of the Society for Simulation in Healthcare*, *20*(3), 158–166. [CrossRef]

Cantrell, M. A. (2008). The importance of debriefing in clinical simulations. *Clinical Simulation in Nursing*, *4*(2), e19–e23. [CrossRef]

Chandler, P., & Sweller, J. (1991). Cognitive load theory and the format of instruction. *Cognition and Instruction*, *8*(4), 293–332. [CrossRef]

Cheng, A., Grant, V., Huffman, J., Burgess, G., Szyld, D., Robinson, T., & Eppich, W. (2017). Coaching the debriefer: Peer coaching to improve debriefing quality in simulation programs. *Simulation in Healthcare*, *12*(5), 319–325. [CrossRef] [PubMed]

Cheng, A., Kolbe, M., Grant, V., Eller, S., Hales, R., Symon, B., Griswold, S., & Eppich, W. (2020). A practical guide to virtual debriefings: Communities of inquiry perspective. *Advances in Simulation*, *5*, 18. [CrossRef]

Chronister, C., & Brown, D. (2012). Comparison of simulation debriefing methods. *Clinical Simulation in Nursing*, *8*(7), e281–e288. [CrossRef]

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334. [CrossRef]

Crookall, D. (2023). *Debriefing: A practical guide*. Springer International Publishing.

Dreifuerst, K. T. (2015). Getting started with debriefing for meaningful learning. *Clinical Simulation in Nursing*, *11*(5), 268–275. [CrossRef]

Dufrene, C., & Young, A. (2014). Successful debriefing—Best methods to achieve positive learning outcomes: A literature review. *Nurse Education Today*, *34*(3), 372–376. [CrossRef]

Eppich, W., & Cheng, A. (2015). Promoting Excellence and Reflective Learning in Simulation (PEARLS): Development and rationale for a blended approach to health care simulation debriefing. *Simulation in Healthcare*, *10*(2), 106–115. [CrossRef]

Evangelou, D., Klar, M., Träg, K., Mulders, M., Marnitz, M., & Rahner, L. (2025, September 8–11). *GenAI-chatbots as debriefers: Investigating the role conformity and learner interaction in counseling training*. 23. Fachtagung Bildungstechnologien (DELFI 2025) (pp. 41–55), Freiberg, Germany. [CrossRef]

Evangelou, D., Mulders, M., & Träg, K. H. (2026). Debriefing in virtual reality simulations for the development of counseling competences: Human-led or AI-guided? *Tech Know Learn*. [CrossRef]

Favolise, M. (2024). Post-simulation debriefing methods: A systematic review. *Archives of Physical Medicine and Rehabilitation*, *105*(4), e146. [CrossRef]

Fey, M. K., & Jenkins, L. S. (2015). Debriefing practices in nursing education programs: Results from a national study. *Nursing Education Perspectives*, *36*(6), 361–366. [CrossRef]

Fraser, K., & McLaughlin, K. (2019). Temporal pattern of emotions and cognitive load during simulation training and debriefing. *Medical Teacher*, *41*(2), 184–189. [CrossRef]

Garden, A. L., Le Fevre, D. M., Waddington, H. L., & Weller, J. M. (2015). Debriefing after simulation-based non-technical skill training in healthcare: A systematic review of effective practice. *Anaesthesia and Intensive Care*, *43*(3), 300–308. [CrossRef] [PubMed]

Grant, J. S., Moss, J., Epps, C., & Watts, P. (2010). Using video-facilitated feedback to improve student performance following high-fidelity simulation. *Clinical Simulation in Nursing*, *6*(5), e177–e184. [CrossRef]

Greer, S. K., Jeffe, D. B., Manga, A., Murray, D. J., & Emke, A. R. (2023). Cognitive load assessment scales in simulation: Validity evidence for a novel measure of cognitive load types. *Simulation in Healthcare: The Journal of the Society for Simulation in Healthcare*, *18*(3), 172–180. [CrossRef] [PubMed]

Hense, J., & Kriz, W. C. (2008). Making simulation games an even more powerful tool. Introducing the theory-based evaluation approach. In L. de Caluwé, G. J. Hofstede, & V. Peters (Eds.), *Why do games work* (pp. 211–217). Kluwer.

Holmes, W., Bialik, M., & Fadel, C. (2019). *Artificial intelligence in education promises and implications for teaching and learning*. Center for Curriculum Redesign. Available online: https://discovery.ucl.ac.uk/id/eprint/10139722/ (accessed on 25 November 2025).

Ifenthaler, D., Majumdar, R., Gorissen, P., Judge, M., Mishra, S., Raffaghelli, J., & Shimada, A. (2024). Artificial intelligence in education: Implications for policymakers, researchers, and practitioners. *Technology, Knowledge and Learning*, *29*(4), 1693–1710. [CrossRef]

INACSL Standards Committee. (2016). INACSL standards of best practice: SimulationSM debriefing. *Clinical Simulation in Nursing*, *12*, S21–S25. [CrossRef]

Kalyuga, S. (2011). Cognitive load theory: How many types of load does it really need? *Educational Psychology Review*, *23*(1), 1–19. [CrossRef]

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., & Hüllermeier, E. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, *103*, 102274. [CrossRef]

Kerlyl, A., Hall, P., & Bull, S. (2007). Bringing chatbots into education: Towards natural language negotiation of open learner Models. In R. Ellis, T. Allen, & A. Tuson (Eds.), *Applications and innovations in intelligent systems XIV* (pp. 179–192). Springer. [CrossRef]

Klar, M. (2025). *Generative AI Chatbots for self-regulated learning while balancing cognitive load: Perceptions, interaction patterns, and instructional designs in K-12 learning* [Doctoral dissertation, Universität Duisburg-Essen].

Klepsch, M., Schmitz, F., & Seufert, T. (2017). Development and validation of two instruments measuring intrinsic, extraneous, and germane cognitive load. *Frontiers in Psychology*, *8*, 1997. [CrossRef]

Kolb, D. A. (2014). *Experiential learning: Experience as the source of learning and development*. FT Press. Available online: https://books.google.com/books?hl=de&lr=&id=jpbeBQAAQBAJ&oi=fnd&pg=PR7&dq=Kolb,+D.+A.+(1984).+Experiential+learning:+Experience+as+the+source+of+learning+and+development.+Englewood+Cliffs,+NJ:+Prentice+Hall.+&ots=Vp6UnY0XRc&sig=yjWgem4qHWdpPjcPPvGRGi_EnXI (accessed on 25 November 2025).

Koole, S., Dornan, T., Aper, L., De Wever, B., Scherpbier, A., Valcke, M., Cohen-Schotanus, J., & Derese, A. (2012). Using video-cases to assess student reflection: Development and validation of an instrument. *BMC Medical Education*, *12*(1), 22. [CrossRef] [PubMed]

Kriz, W. C., & Nöbauer, B. (2015). *Den lernerfolg mit debriefing von Planspielen sichern*. Bertelsmann.

Kriz, W. C., Saam, N., Pichlbauer, M., & Fröhlich, W. (2007). Intervention mit planspielenals Großgruppenmethode–Ergebnisse einer interviewstudie. In *Planspiele für die organisationsentwicklung. schriftenreihe: Wandel und kontinuität in organisationen* (Vol. 8, pp. 103–122). Wissenschaftlicher Verlag.

Kumar, P., Harrison, N. M., McAleer, K., Khan, I., & Somerville, S. G. (2025). Exploring the role of self-led debriefings within simulation-based education: Time to challenge the status quo? *Advances in Simulation*, *10*(1), 9. [CrossRef]

Leppink, J., Paas, F., Van der Vleuten, C. P., Van Gog, T., & Van Merriënboer, J. J. (2013). Development of an instrument for measuring different types of cognitive load. *Behavior Research Methods*, *45*(4), 1058–1072. [CrossRef]

Liang, J.-C., & Hwang, G.-J. (2023). A robot-based digital storytelling approach to enhancing EFL learners' multimodal storytelling ability and narrative engagement. *Computers & Education*, *201*, 104827. [CrossRef]

Luctkar-Flude, M., Tyerman, J., Verkuyl, M., Goldsworthy, S., Harder, N., Wilson-Keates, B., Kruizinga, J., & Gumapac, N. (2021). Effectiveness of debriefing methods for virtual simulation: A systematic review. *Clinical Simulation in Nursing*, *57*, 18–30. [CrossRef]

McDonald, G. (2013). Does size matter? The impact of student–staff ratios. *Journal of Higher Education Policy and Management*, *35*(6), 652–667. [CrossRef]

Memarian, B., & Doleck, T. (2023). ChatGPT in education: Methods, potentials, and limitations. *Computers in Human Behavior: Artificial Humans*, *1*(2), 100022. [CrossRef]

Metcalfe, S. E., Hall, V. P., & Carpenter, A. (2007). Promoting collaboration in nursing education: The development of a regional simulation laboratory. *Journal of Professional Nursing*, *23*(3), 180–183. [CrossRef]

Miller, C. R., Greer, S. K., Toy, S., & Schiavi, A. (2025). Debriefing is germane to simulation-based learning: Parsing cognitive load components and the effect of debriefing. *Simulation in Healthcare: The Journal of the Society for Simulation in Healthcare*, *20*(6), 349–356. [CrossRef]

Na, Y. H., & Roh, Y. S. (2021). Effects of peer-led debriefing on cognitive load, achievement emotions, and nursing performance. *Clinical Simulation in Nursing*, *55*, 1–9. [CrossRef]

Nghi, T. T., & Anh, L. T. Q. (2024). Promoting student-centered learning strategies via AI chatbot feedback and support: A case study at a public university in Vietnam. *International Journal of Teacher Education and Professional Development (IJTEPD)*, *7*(1), 1–25. [CrossRef]

Ortega-Ochoa, E., Arguedas, M., & Daradoumis, T. (2024). Empathic pedagogical conversational agents: A systematic literature review. *British Journal of Educational Technology*, *55*(3), 886–909. [CrossRef]

Paas, F., Renkl, A., & Sweller, J. (2003). Cognitive load theory and instructional design: Recent developments. *Educational Psychologist*, *38*(1), 1–4. [CrossRef]

Palaganas, J. C., Fey, M., & Simon, R. (2016). Structured debriefing in simulation-based education. *AACN Advanced Critical Care*, *27*(1), 78–85. [CrossRef]

Rudolph, J. W., Simon, R., Dufresne, R. L., & Raemer, D. B. (2006). There's no such thing as "nonjudgmental" debriefing: A theory and method for debriefing with good judgment. *Simulation in healthcare*, *1*(1), 49–55. [CrossRef] [PubMed]

Ryoo, E. N., & Ha, E. H. (2015). The importance of debriefing in simulation-based learning: Comparison between debriefing and no debriefing. *CIN: Computers, Informatics, Nursing*, *33*(12), 538–545. [CrossRef] [PubMed]

Sawyer, T., Eppich, W., Brett-Fleegler, M., Grant, V., & Cheng, A. (2016). More than one way to debrief: A critical review of healthcare simulation debriefing methods. *Simulation in Healthcare*, *11*(3), 209–217. [CrossRef] [PubMed]

Shinnick, M. A., Woo, M., Horwich, T. B., & Steadman, R. (2011). Debriefing: The most important component in simulation? *Clinical Simulation in Nursing*, *7*(3), e105–e111. [CrossRef]

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, *12*(2), 257–285. [CrossRef]

Sweller, J. (2010). Cognitive Load Theory: Recent Theoretical Advances. In J. L. Plass, R. Moreno, & R. Brünken (Eds.), *Cognitive load theory* (pp. 29–47). Cambridge University Press.

Sweller, J. (2011). Cognitive load theory. In *Psychology of learning and motivation* (Vol. 55, pp. 37–76). Academic Press.

Sweller, J., Van Merriënboer, J. J. G., & Paas, F. (2019). Cognitive architecture and instructional design: 20 years later. *Educational Psychology Review*, *31*(2), 261–292. [CrossRef]

Tosterud, R., Hedelin, B., & Hall-Lord, M. L. (2013). Nursing students' perceptions of high-and low-fidelity simulation used as learning methods. *Nurse Education in Practice*, *13*(4), 262–270. [CrossRef]

Wang, M., & Akhter, S. (2025). Tracing interpersonal emotion regulation, behavioral emotion regulation strategies, hopelessness and vocabulary retention within Bing vs. ChatGPT environments. *British Educational Research Journal*, 1–28. [CrossRef]

Winkler, R., & Soellner, M. (2018). Unleashing the potential of chatbots in education: A state-of-the-art analysis. *Academy of Management Proceedings*, *2018*(1), 15903. [CrossRef]

Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education—Where are the educators? *International Journal of Educational Technology in Higher Education*, *16*(1), 39. [CrossRef]

Zhu, X. T., Cheerman, H., Cheng, M., Kiami, S. R., Chukoskie, L., & McGivney, E. (2025, April 26–May 1). *Designing VR simulation system for clinical communication training with LLMs-based embodied conversational agents*. Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (pp. 1–9), Okohama, Japan. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

https://doi.org/10.3390/educsci16020255