

# GenAI-Chatbots as Debriefers: Investigating the Role Conformity and Learner Interaction in Counseling Training

Dominik Evangelou <sup>1</sup>, Maria Klar <sup>1</sup>, Kristian Träg <sup>1</sup>, Miriam Mulders <sup>1</sup>, Melina Marnitz <sup>1</sup>, and Lara Rahner <sup>6</sup>

**Abstract:** Debriefing is essential for the effectiveness of simulation-based training but is generally considered resource-intensive. Generative Artificial Intelligence (GenAI)-based debriefing can be an alternative to human debriefers. However, there is no research yet whether GenAI chatbots can take up this role and how learners react to them. This paper presents a qualitative analysis of a debriefing following a counseling training conducted in Virtual Reality (VR) with the support of a GenAI chatbot. The debriefing helped students to analyze their experiences and application of consulting techniques. The analysis of the chatlogs (n = 22) are focused on the role conformity of the bot and the students' ability to reflect their behavior within VR. The results revealed the chatbot's strong role conformity but also its tendency for overly complimentary answers. However, this bias does not seem to influence students' self-reflection. Instead, they maintained a self-critical attitude. Future research on AI-assisted debriefing could expand on these findings in related areas.






**Keywords:** Counseling, Virtual Reality, Training, Chatbot, AI, GenAI, Debriefing, Roleconformity, Student behavior




## 1 Introduction

Effective learning through simulations and serious games critically depends on the structured processing and transformation of experiences in subsequent debriefing sessions [Cr23]. The pedagogical value of even the most technologically and aesthetically advanced simulations can only be fully realized when learners are given the opportunity to reflect systematically on their actions and decisions. As argued by [Th90], such reflective engagement serves as a central mechanism for integrating experience into meaningful knowledge structures and facilitating transfer to novel contexts.

Despite its recognized importance, debriefing—particularly in Virtual Reality (VR) and simulation-based learning environments—remains a resource-intensive practice, often constrained by the limited availability of trained facilitators [MHC07]. While research on simulation-based debriefing is steadily growing, it remains fragmented across domains, with

<sup>1</sup> University of Duisburg-Essen, Chair of Educational Technology and Instructional Design, Universitätsstraße 2, 45117 Essen, Germany, dominik.evangelou@uni-due.de,  <https://orcid.org/0009-0002-7152-1594>; maria.klar@uni-due.de,  <https://orcid.org/0000-0002-5298-8458>; kristian.traeg@uni-due.de,  <https://orcid.org/0009-0004-4557-2026>; miriam.mulders@uni-due.de,  <https://orcid.org/0000-0003-0683-2310>; melina.marnitz@rwth-aachen.de,  <https://orcid.org/0009-0007-8373-4729>

<sup>6</sup> TU Dortmund, Center for Higher Education, Hohe Straße 141, 44139 Dortmund, Germany, lara.rahner@tu-dortmund.de,  <https://orcid.org/0009-0000-1298-8640>

the healthcare sector currently at the forefront of empirical and conceptual developments (e. g., [DY14; Fa24; Ga15; Lu21]).

Recent advancements in generative artificial intelligence (GenAI), especially in the form of large language model (LLM)-based chatbots, offer new opportunities to support reflective learning processes at scale. These systems may be capable of simulating debriefing dialogues that approximate human-facilitated reflection. The present study investigates this potential by examining the chatbot's ability to maintain role conformity in the function of a debriefer, as well as students' responses to and engagement with such AI-mediated interactions.

## **2 Background**

### **2.1 Debriefing**

The integration of a debriefing following the actual learning situation has long been an essential component in educational contexts, even before the availability of simulations in VR. Research on simulation games and role-playing has already shown that meaningful learning effects cannot be achieved merely by exposing learners to a model of reality that is both appropriately complex and authentic. Rather, in addition to providing complex learning scenarios, simulation game instructional frameworks – such as the design of an effective debriefing for reflection and learning transfer – are necessary to transfer the experiences from the simulation games to real-world professional practice [KS07]. The simulation game method is particularly well-suited for developing competencies, as it applies the principles of problem-based learning. Learners are confronted with complex and authentic situations that enable experiential learning. It is essential that learners engage in cooperative learning through team-based action and receive appropriate support from the instructor – including in the form of joint reflection. Notably, the aspect of collective reflection directly relates to the debriefing process. Accordingly, [KS07] emphasizes that only knowledge acquired through reflection can be transferred to unfamiliar domains. It follows that especially when learning through online courses or acquiring competencies within virtual reality simulations, the content to be learned needs to be reviewed and discussed in a debriefing afterward.

Debriefing can generally be conducted in two fundamentally different ways: as a guided reflection or as a self-debriefing. Additionally, these two methods can be varied by factors such as the number of participants, the instructions, technical aids, or even the instructor themselves (e. g. [DY14; Fa24; Lu21]). In guided debriefing, the instructor moderates the process by posing questions related to the pre-defined learning objectives. Particularly for inexperienced learner groups, facilitating the reflection process seems to be of great importance [KS07]. According to [Th11], this type of debriefing is considered the most suitable reflection method. A review of literature reveals that guided debriefing is indeed the most frequently used method (see [Bo11; Ti13; TTH15; Ve18]); however, there is no empirical evidence indicating that this method is also the most effective [DY14]. In contrast,

unguided debriefing allows learners to reflect on their learning experiences independently and self-organized, for example, within a group setting. Studies show that there is no statistically significant difference between guided debriefings and self-debriefings. Both methods lead to performance improvements [Bo11; Bo13].

In addition to comparing guided and unguided debriefings, several studies have examined variations in specific debriefing aspects. For instance, some studies compare the effectiveness of asynchronous debriefings [At21], or video-assisted debriefing [Gr10]. These studies further confirm the finding that any form of debriefing has positive effects on performance. The only factor that appears to lead to significant differences in debriefing quality is the instructor's skill as a debriefer [Ti13; TTH15]. Studies focusing on the debriefing of simulations are, as previously mentioned, primarily found in the healthcare sector. A literature review by [DY14] indicates that research on post-simulation debriefings is still in its infancy. The researchers identified only 13 articles that either compared various debriefing strategies or examined student perceptions of debriefing. The relatively limited amount of relevant research literature is further supported by the reviews of [Ga15] and [Lu21], who describe eight and seven articles, respectively, comparing different debriefing strategies or explored student perceptions of debriefing.

It can be stated that post-simulation debriefing plays a crucial role in learning. Research indicates that debriefing can be resource-intensive, depending on the format, and that no single method has been proven more effective than others to date [DY14]. This highlights the significant potential for further exploration of different debriefing methods. The following sections will explore the potential use of AI in the context of debriefing. The focus will first be on GenAI Chatbots, followed by an examination of their potential application as debriefers.

## 2.2 Chatbots

While scripted Chatbots have been around for decades, AI-based generative Chatbots have taken off only recently, notably with the publication of ChatGPT in late 2022 [Wa24]. These types of chatbots are large language models (LLMs), meaning they have been fed large amounts of text-data to analyze the likelihood of words occurring together, and use those likelihoods to be able to generate new coherent sentences [Na23].

Researchers from varying fields have investigated the use of GenAI-chatbots in education. [Ho23] identifies authentic language use as a key benefit that GenAI offers for language teaching, in addition to personalized tutoring via AI. [CH23] find in a survey that university students in Hong Kong hold generally positive attitudes towards GenAI in teaching and learning and would like to integrate them into their learning practices, for example for personalized feedback. However, many students seemed concerned about topics like privacy, ethics, and transparency when handling GenAI, as well as over-reliance on AI-technology and the degradation of skills accompanied by it [CH23]. [SQ24] show that texts that ChatGPT

generated for potential use in climate change education were scientifically accurate, but did not address certain economic activities and actors that contribute to increased carbon emissions. Overall, educators seem to have mixed experiences with GenAI, with tertiary education being the main focus of research up to this point [JLC23; Kh22]. The present paper attempts to expand upon this body of research by introducing a GenAI chatbot as a conversational partner to training for counsellors.

Utilizing GenAI chatbots as conversational partners in education might lead to a few problems. For educators, it might be difficult to use and teach a technology that they themselves are still unfamiliar with [Ng23]. For learners, the heavy amount of cognitive offloading afforded by assisting chatbots may lead to a lack of own contributions, possibly resulting in placebo effects when learning [Sk24]. The following section will go into more depth on using chatbots during debriefings of training interventions where these challenges need to be kept in mind.

### **2.3 Chatbots as Debriefers**

As stated earlier, debriefer skills are essential for debriefing quality [Ti13; TTH15], so the question is whether current GenAI chatbots are performant enough to fill in the debriefer role. Debriefing requires a dynamic interaction between the debriefer and the learner that relates back to the VR simulation. The debriefer's role, therefore, requires going on tangents in a conversation and coming back to maintain the overall thread of conversation. While LLMs can be used to generative one-off texts like feedback [Da23] or test questions [Le24], it is more of a challenge to set up LLMs in such a way that they can hold up a complex conversation over a longer period [Ji24]. For AI-led debriefing the question arises whether role adherence is good enough to reach a sufficient quality.

Sycophancy, an LLM's tendency to generate output that agrees with or pleases the user, could be a further issue in the context of debriefing. This phenomenon might have adverse effects on perceived authenticity and trust [SW25]. Authenticity and trust are key factors not only in successful counselling, but also for debriefings [Ro57; Ro62; SFE16]. As debriefing requires the critical reflection of the learner's experience and actions, it would not be beneficial to the learning process to have an overly positive and uncritical chatbot debriefer [Ma24]. It is therefore vital to evaluate chatbot debriefers for this tendency.

However, even if the chatbot did not adhere to the role perfectly or displayed sycophantic or other awkward behavior (e. g., overly praising the learner or breaking the role), this would not necessarily mean it could not be used for debriefing as learners themselves could mitigate these weaknesses to some degree [Ji24]. The evidence that self-guided debriefing can be equally effective indicates that learners can take on responsibility for the debriefing process and that the quality of the debriefing does not only depend on the debriefer (human or AI) but also the learner. The responsibility for the debriefing process can be shared

between learner and AI [JNH23; Mo22]. Nonetheless, there is little research on how learners interact with GenAI chatbots and whether learners can amend chatbot weaknesses.

There is research showing that learners have difficulty to prompt chatbots effectively [Za23] and that they do not use chatbots as a dialogue partner but rather an information source [K124]. When students interacted with a chatbot versus a peer in problem-solving task, it was found that they interacted less dialogically and more predictably with the chatbot than with the peer, but preferred the chatbot nevertheless [So25]. Based on these findings, it could be expected that students do not easily engage in a fluent debriefing conversation and amend chatbot weaknesses.

## 2.4 Research Questions

In this article, we aim to investigate the extent to which GenAI chatbots can be utilized for post-simulation debriefings in educational contexts and thus represent a resource-saving but still significant alternative to classic debriefings. When analyzing the chatlogs, we focus on the aspects of debriefers' role conformity and student behavior: (1) Does a GenAI chatbot maintain its role as a debriefer? (2) How do students interact with the chatbot debriefer?

## 3 Methods

### 3.1 Procedure

The study was conducted as part of a VR training aimed at enhancing counseling skills among participants. The training session took place in a controlled lab setting, where participants engaged in a simulated counseling interaction with a standardized client (avatar) within a VR environment. Prior to the training, participants were informed about the study and asked to provide informed consent. They were then assigned a unique pseudonym code to ensure anonymization throughout the study.

After completing a pre-questionnaire, participants received instructions on operating the VR equipment and were introduced to the counseling case. They then entered the VR environment using a head-mounted display and interacted with the virtual client for approximately 20 minutes. Participants applied previously learned techniques such as summarizing and asking open-ended questions during the consultation.

Following the VR session, participants completed a second questionnaire. Subsequently, they took part in one of two types of debriefing: either a guided debriefing led by a trained facilitator or a structured, GenAI chatbot-based debriefing. As we are only referring to chatbot-based debriefing in this article, we will only describe this in the following. The chatbot's task was to guide the test participants through the debriefing. The chatbot should

follow a pre-coded protocol aligned with the three-phase debriefing model [PFS16]. To enable authentic communication with a role-conforming chatbot, we tested several system prompts in advance. In doing so, we tested many slightly varied prompts and used different LLMs. We optimized our prompt in iterative process loops with several research assistants and student participants. The system prompt employed in this study is documented in the online appendix [Ev25]. For the underlying model, we selected Meta Llama 3.1 8B Instruct, as it demonstrated robust performance despite its relatively small parameter size. Temperature and nucleus sampling were both set to 0.5 to enable sampling beyond the most common word while retaining some level of coherence in the chatbot's responses. The study concluded with a final post-questionnaire and informing the participants about the study objectives.

To evaluate the effectiveness of the training, participants' self-efficacy and counseling competence were assessed at all measurement time points. These constructs were selected as key indicators of participants' confidence in their own counseling abilities and their perceived skill development through the VR-based learning experience. Self-efficacy and counseling competence were measured using scales based on [He09]. In addition, the third measurement time point also included participants' perception of the debriefing experience, which was assessed using the Debriefing Experience Scale (DES) developed by [Re12].

The participants were able to start the debriefing with the chatbot themselves. They used a laptop provided by us for this purpose. Next to them was a written guide with instructions (e. g., "Be as detailed as possible and give specific examples."), which they had to follow during the 20-minutes conversation.<sup>7</sup> The study concluded with a final post-questionnaire and informing the participants about the study objectives.

To evaluate the effectiveness of the training, participants' self-efficacy and counseling competence were assessed at all measurement time points. These constructs were selected as a key indicators of participants' confidence in their own counseling abilities and their perceived skill development through the VR-based learning experience. Self-efficacy and counseling competence were measured using scales based on [He09]. In addition, the third measurement time point also included participants' perception of the debriefing experience, which was assessed using the *Debriefing Experience Scale* developed by [Re12].

### 3.2 Participants

Participants were students at a large university in Germany enrolled in a higher education counseling training seminar. All participants had prior exposure to theoretical knowledge of counseling techniques. The final sample consisted of 46 persons, of which 22 were randomly assigned to the chatbot condition. Of these 22 (age  $M = 23.3$ ;  $SD = 4.3$ ), 20 were female and 2 were male.

---

<sup>7</sup> The full guide can be accessed via [Ev25].

### 3.3 Data Analysis

The chatlogs of the 22 dialogs were saved and analyzed according to the principles of qualitative content analysis [KR20]. Two distinct coding schemes were developed by two experienced scientists for this purpose, one relating to research question 1, i. e., role conformity, and another one relating to research question 2, i. e., student behavior. Both coding schemes were applied using MAXQDA. The first coding scheme assessed whether and how the chatbot adhered to its predefined role as a debriefer, based on the three-phase model of debriefing by [PFS16]. Furthermore, this coding scheme captured structural elements of the chatbot's language and behavior. The second coding scheme assessed how participants reflected on their counseling experience and the debriefing process. All codes were inductively developed based on recurring themes in the participants' written responses. The tables each contain a brief description of the (sub-)code, one or more direct quotes from the chatlogs for each code, and the cumulative number of mentions across all chatlogs.<sup>8</sup>

After the study, two experienced research assistants independently examined the chatlogs according to the (sub)-codes contained in the coding schemes. More than a third ( $n = 8$ ) were double-coded by the assistants. The interrater reliability of  $\kappa = .82$  is in the acceptable range [Co68; Ku19].

We supplemented our qualitative data from chatlogs with one item of the Debriefing Scale ([Re12] from the final questionnaire, in which students were asked to rate their agreement to the statement "The chatbot provided adequate guidance during the debriefing" on a Likert-scale from 1 (*totally disagree*) to 5 (*totally agree*).

## 4 Results

The following results section presents the qualitative analysis of the chatlogs, focusing on the chatbot's role conformity and students' behavior. A total of eighteen categories, including subcategories, were identified for role conformity, and fourteen categories for students' behavior, using both deductive and inductive approaches. Due to the large number of categories, the complete tables containing all codes and exemplary quotes for each code are provided in the online appendix. Tables 1 and 2 are intended to illustrate exemplary and interesting results of the qualitative analysis.

### 4.1 Chatbot's Role conformity

In terms of role conformity, the chatbot adhered to most of the required steps outlined in the system prompt. In all but one chat, it introduced itself as the debriefer. Notably, in one

<sup>8</sup> The complete tables with codes and subcodes can be found at [Ev25].

case the participant did not enter the designated keyword “start” to initiate the debriefing as recommended. This likely caused the opening of the debriefing to deviate from the system prompt, as seen in the example: *“Thank you for being willing to speak with me. I hear that you are feeling [. . .] a bit overwhelmed after completing the VR training.”* Interestingly, despite this deviation, the chatbot proceeded with the debriefing according to the three-phase structure. Despite the presumably less authentic communication, this student rated the chatbot-guided debriefing as highly appropriate, giving it a 5 out of 5. The other participants also provided consistently positive ratings ( $M = 4.33$ ,  $SD = 0.58$ ). After its introduction, the chatbot outlined the debriefing process in all cases and guided the conversation through the respective phases using appropriate transitions. Remarkably, only in transcript SD10 did it explicitly conclude the reaction phase with the statement: *“This is the conclusion of the reaction phase.”* In contrast, transitions to the understanding phase ( $\Sigma = 21$ ; 16/22) – the symbol  $\Sigma$  represents the total number of mentions across all chat transcripts, while the fraction (x/22) indicates the number of chat transcripts in which this specific (sub)code appears – and summary phase ( $\Sigma = 14$ ; 14/22) occurred more frequently. As the conversation progressed, the chatbot’s outputs primarily focused on prompting reflection on the learning content ( $\Sigma = 236$ ; 22/22). For example, it asked: *“We have learned three techniques: summarizing, paraphrasing, mirroring, I-messages, and questioning techniques. Which technique did you find most difficult or most interesting?”* In addition to reflecting on conversation techniques, applying learned content to real-world practice was a key objective of the debriefing. The chatbot facilitated this transfer ( $\Sigma = 86$ ; 19/22) by posing targeted questions such as: *“Which aspects do you think are most important, and how can you integrate them into your counseling practice?”* It is also noteworthy that the chatbot actively addressed students’ emotions. This was achieved, in part, through direct inquiries about their emotional state, as illustrated in table 1. In fact, all chatlogs included at least one such inquiry. Moreover, in sixteen transcripts ( $\Sigma = 39$ ) the chatbot demonstrated attempts to convey understanding or empathy in response to students’ input (see Table 1). In the final phase of the debriefing, students were expected to summarize their key takeaways. The chatbot frequently facilitated this process by posing targeted questions such as: *“What are the key insights you can take away from this debriefing?”* In some instances, it also generated independent summaries of students’ responses (see Table 1). Across eighteen transcripts, the summarization phase was explicitly referenced 38 times.

Beyond the content-driven aspects of the chatbot debriefer, which were shaped by the system prompt, certain chatbot-specific characteristics may have contributed to a perception of inauthentic communication. For instance, twelve transcripts ( $\Sigma = 28$ ) contain stage directions, such as: *“(Pause to give the student time to collect their thoughts and mention key points.)”* These instructions typically appeared after the chatbot posed a question. Authenticity in conversation can also be disrupted by the chatbot’s tendency toward sycophantic behavior, as discussed in Section 2.3. This is particularly evident in its overly enthusiastic responses, which appear in every chatlog. For example, it frequently used adjectives such as interesting ( $\Sigma = 30$ ; 13/22), awesome ( $\Sigma = 67$ ; 19/22), fantastic ( $\Sigma = 5$ ; 3/22), and very good ( $\Sigma = 40$ ; 13/22). As shown in table 1, the word important appears 156 times across 21 debriefings.



Codes and Subcodes	Examples
Asks about emotions ( $\Sigma = 35$ ; 22/22)	<i>Let's dive deeper into your feelings. How did you feel when you began to understand the situation better and started applying the counseling techniques? Did you feel more confident, or were there still uncertainties?</i>
Expresses empathy ( $\Sigma = 39$ ; 16/22)	<i>It is understandable that you didn't think about taking notes in that situation. The situation was complex, and you were focused on communicating with the client.</i>
Summary of key findings ( $\Sigma = 38$ ; 18/22)	<i>Let's now summarize what we have discussed:</i> <ul style="list-style-type: none"> <li>• <i>You felt comfortable while conducting the VR training.</i></li> <li>• <i>[...]</i></li> </ul>
Uses adjective "important" ( $\Sigma = 156$ ; 21/22)	<i>That is an important step toward becoming a successful counselor.</i>

Tab. 1: Exemplary codes for role conformity ( $\Sigma$  = number of mentions across chatlog files; x/22 = share of chatlog files with this (sub)code)

Additionally, the chatbot exhibited repetition within individual conversations. For instance, the sentence: *"That shows that you engaged with the situation and that you were willing to open up and try new things."* appears 16 times in one case.

## 4.2 Student behavior

Next to the chatbot behavior, the question is how students reacted to the chatbot and whether they engaged in a reflection that fulfill the function of a debriefing. The main function of a debriefing is a critical review of the simulation to facilitate transfer to real-world scenarios. Furthermore, a debriefing should allow for expression of affective responses to the simulation, especially if the simulation was emotionally activating [Cr23].

In this study, most students indeed processed their current emotional state once or several times (see Table 2). They also evaluated their role as a counselor either positively, neutrally, or negatively at least once (see Table 2). In absolute numbers, they assessed their performance positively the least often, but in relative numbers negative self-appraisal occurred the least often. Assuming a normal distribution of actual counseling performance, these reflections can be seen as adequately balanced. There were a considerable number of statements on transfer to real-world counseling (see Table 2). Similarly, 14 of the 22 participants felt that they have learned something in the simulation ( $\Sigma = 20$ ). This is in line with a rather large number of assessments of the counseling techniques ( $\Sigma = 36$ ; 16/22), e. g., *"Yes, the*

*questioning techniques helped me better understand the person's concerns and structure the conversation accordingly*", as well as a comparatively large number of statements on the difficulty level ( $\Sigma = 14$ ; 10/22), e. g., "[. . . ] I concentrated more on paraphrasing and mirroring. I found the I-Messages difficult to implement. [. . . ] I was unsure whether they were the right questions [. . . ]."

Next to these reflections on the content level, there were a number of reflections on the meta level. 17 participants ( $\Sigma = 27$ ) chatted about the simulation and study situation, e. g., *"I think one challenge of VR training is that facial expressions and general body language are currently difficult to represent in the digital space. [. . . ] However, the VR interface allows for greater anonymity [. . . ]."* More than half expressed ideas on what could help or could have helped them in their learning process: *"Observing an experienced counselor and focusing on how they apply techniques could be helpful [. . . ]."* One student discussed the debriefing itself three times, e. g., by writing: *"The debriefing forces me [. . . ] to reflect on my perceptions and consider whether the counseling was successful or not. It helps me clearly understand what happened in the conversation and realize if and how I applied different techniques"* Only rarely did the students express that they do not know how to answer ( $\Sigma = 6$ ; 6/22), use conversational phrases like *"thank you"* ( $\Sigma = 2$ ; 2/22), and only in one case did the student take the lead three times and moderated the conversation, e. g., *"Instead of continuing to discuss the person's situation, I would like to talk about how I can train my counseling skills in the future."*

Codes and Subcodes	Examples
Current emotional state of the participant ( $\Sigma = 35$ ; 20/22)	<i>I was a bit nervous because I had never done anything like this before. The situation was unfamiliar [. . . ] I wasn't sure what to do.</i>
Positive reflection on own behavior ( $\Sigma = 16$ ; 13/22)	<i>Still, I am proud of myself for conducting a counseling conversation, and I am mostly satisfied.</i>
Neutral reflection of own behavior ( $\Sigma = 24$ ; 14/22)	<i>I mainly used open-ended questions because I wanted to learn as much as possible about the client's feelings and thoughts. I also used I-Messages to show that I appreciated the fact that the person already had certain goals in mind.</i>
Critical reflection on own behavior ( $\Sigma = 24$ ; 10/22)	<i>[. . . ] I quickly chose the Tree-of-Life method, even though there might have been better alternatives.</i>

Tab. 2: Exemplary codes for student behavior ( $\Sigma$  = number of mentions across chatlog files; x/22 = share of chatlog files with this (sub)code)

## 5 Discussion

Our findings indicate that the GenAI chatbot predominantly fulfilled its function as a debriefer in accordance with the structured three-phase debriefing model. A key limitation was its tendency toward an overly enthusiastic tone and insufficient critical reflection, consistent with prior observations on chatbot sycophancy [Ma24]. Notably, students appeared unaffected by this bias, maintaining critical self-reflection, which aligns with evidence highlighting learner agency in self-guided debriefing [VMS21].

Despite largely conforming to its assigned role, the chatbot exhibited deviations, particularly when interaction patterns diverged from expectations (e. g., participant JE21). It occasionally breached the “fourth wall” by inserting meta-communicative stage directions (e. g., “waiting for the learner to respond”) that potentially disrupted conversational flow [Ka25]. These occurrences may be linked to system parameters—temperature and nucleus sampling fixed at 0.5—and warrant further investigation for optimal configuration to mitigate such behavior.

The probabilistic nature of AI-generated responses complicates consistent conversational quality, as evidenced by limited dynamic engagement and repetitive phrasing [Ba24; Ji24]. Nevertheless, participants did not express dissatisfaction with the chatbot’s output and appeared to accommodate its limitations, as reflected by favorable guidance ratings. This suggests that learners’ prior awareness of large language model constraints may facilitate adaptive interaction strategies.

From a methodological perspective, employing a relatively small LLM likely contributed to observed limitations. Larger models generally offer enhanced contextual understanding and reduced sycophancy, representing a promising avenue for future research. However, given that many educational institutions have greater access to small- or medium-scale models, our findings demonstrate that complex instructional tasks such as debriefings can be effectively supported by smaller architectures. Future work should systematically evaluate alternative system prompts aimed at curbing exaggerated or overly agreeable responses within structured debriefing protocols.

Critically, this study was conducted with experimenter presence. Subsequent investigations should examine fully autonomous chatbot-based debriefings, which hold promise for resource-efficient and scalable applications, particularly in MOOCs or online training environments with limited human facilitation.

Finally, as participants were university students with foundational knowledge of counseling techniques, generalizability to more diverse or professional populations remains uncertain. Further research should assess differential learner interactions with AI debriefers across varying expertise levels and domains.

In conclusion, notwithstanding occasional role deviations, sycophantic tendencies, and repetitive output, GenAI chatbots demonstrate potential as debriefing facilitators. Learners’

capacity to compensate for chatbot shortcomings underscores their viability as scalable alternatives to human debriefers.

## 6 Conclusion

The aim of this qualitative study was to examine whether generative chatbots are capable of adhering to their assigned role as debriefers within more complex conversational settings, and how learners interact with them. The results indicate that chatbots are capable of conducting reflective conversations in the role of a debriefer, thereby supporting learners in reflecting on educational content.

The relevance of this study becomes evident in light of the resource-intensive nature of guided debriefings conducted by instructors, as recommended by the International Nursing Association for Clinical Simulation and Learning [Co16]. In higher education in particular, the personnel effort required to conduct individual debriefings is often unsustainable [Ch17]. Our findings support the notion that chatbots can effectively guide students through reflective processes. As such, AI-supported debriefings may serve as a resource-efficient complement to moderated debriefing in (higher) education.

Building on these findings, future research should focus especially on the quantitative evaluation of chatbot debriefings and systematically compare their effectiveness with human-led debriefing sessions.

## References

- [At21] Atthill, S. et al.: Exploring the Impact of a Virtual Asynchronous Debriefing Method after a Virtual Simulation Game to Support Clinical Decision-Making. en, *Clinical Simulation in Nursing* 50, pp. 10–18, 2021, DOI: 10.1016/j.ecns.2020.06.008.
- [Ba24] Bai, Y. et al.: Investigating the Efficacy of ChatGPT-3.5 for Tutoring in Chinese Elementary Education Settings. *IEEE Transactions on Learning Technologies* 17, pp. 2102–2117, 2024, DOI: 10.1109/TLT.2024.3464560.
- [Bo11] Boet, S. et al.: Looking in the mirror: Self-debriefing versus instructor debriefing for simulated crises\*: en, *Critical Care Medicine* 39 (6), pp. 1377–1381, 2011, DOI: 10.1097/CCM.0b013e31820eb8be.
- [Bo13] Boet, S. et al.: Within-Team Debriefing Versus Instructor-Led Debriefing for Simulation-Based Education: A Randomized Controlled Trial. en, *Annals of Surgery* 258 (1), pp. 53–58, 2013, DOI: 10.1097/SLA.0b013e31829659e4.
- [Ch17] Cheng, A. et al.: Coaching the debriefer: peer coaching to improve debriefing quality in simulation programs. *Simulation in Healthcare* 12 (5), pp. 319–325, 2017.
- [CH23] Chan, C. K. Y.; Hu, W.: Students' voices on generative AI: perceptions, benefits, and challenges in higher education. *International Journal of Educational Technology in Higher Education* 20 (1), p. 43, 2023, DOI: 10.1186/s41239-023-00411-8.

- [Co16] Committee, I. S. et al.: INACSL standards of best practice: SimulationSM debriefing. *Clinical Simulation in Nursing* 12, S21–S25, 2016.
- [Co68] Cohen, J.: Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin* 70 (4), Place: US Publisher: American Psychological Association, pp. 213–220, 1968, DOI: 10.1037/h0026256.
- [Cr23] Crookall, D.: Debriefing: A Practical Guide. en, *Simulation for Participatory Education*, 2023.
- [Da23] Dai, W. et al.: Can Large Language Models Provide Feedback to Students? A Case Study on ChatGPT. In: 2023 IEEE International Conference on Advanced Learning Technologies (ICALT). IEEE, Orem, UT, USA, pp. 323–325, 2023, DOI: 10.1109/ICALT58122.2023.00100.
- [DY14] Dufrene, C.; Young, A.: Successful debriefing — Best methods to achieve positive learning outcomes: A literature review. en, *Nurse Education Today* 34 (3), pp. 372–376, 2014, DOI: 10.1016/j.nedt.2013.06.026.
- [Ev25] Evangelou, D.: Online Supplement\_Chatsbots as Debriefers, 2025, <https://osf.io/qe5xv>.
- [Fa24] Favolise, M.: Post-Simulation Debriefing Methods: A Systematic Review. en, *Archives of Physical Medicine and Rehabilitation* 105 (4), e146, 2024, DOI: 10.1016/j.apmr.2024.02.680.
- [Ga15] Garden, A. L. et al.: Debriefing after Simulation-Based Non-Technical Skill Training in Healthcare: A Systematic Review of Effective Practice. en, *Anaesthesia and Intensive Care* 43 (3), pp. 300–308, 2015, DOI: 10.1177/0310057X1504300303.
- [Gr10] Grant, J. S. et al.: Using Video-Facilitated Feedback to Improve Student Performance Following High-Fidelity Simulation. en, *Clinical Simulation in Nursing* 6 (5), e177–e184, 2010, DOI: 10.1016/j.ecns.2009.09.001.
- [He09] Hertel, S.: *Beratungskompetenz von Lehrern*. Waxmann Verlag, 2009.
- [Ho23] Hong, W. C. H.: The impact of ChatGPT on foreign language teaching and learning: opportunities in education and research. en, *Journal of Educational Technology and Innovation* 5 (1), Number: 1, pp. 37–45, 2023, <https://jeti.thewsu.org/index.php/ciet/article/view/103>.
- [Ji24] Jin, H. et al.: Teach AI How to Code: Using Large Language Models as Teachable Agents for Programming Education. In: *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. CHI '24, Number: 652, Association for Computing Machinery, New York, NY, USA, pp. 1–28, 2024, DOI: 10.1145/3613904.3642349.
- [JLC23] Jeon, J.; Lee, S.; Choi, S.: A systematic review of research on speech-recognition chatbots for language learning: Implications for future directions in the era of large language models. *Interactive Learning Environments* 0 (0), pp. 1–19, 2023, DOI: 10.1080/10494820.2023.2204343.
- [JNH23] Järvelä, S.; Nguyen, A.; Hadwin, A.: Human and artificial intelligence collaboration for socially shared regulation in learning. en, *British Journal of Educational Technology* 54 (5), pp. 1057–1076, 2023, DOI: 10.1111/bjet.13325.
- [Ka25] Kaate, I. et al.: The ‘fourth wall’ and other usability issues in AI-generated personas: comparing chat-based and profile personas. en, *Behaviour & Information Technology*, pp. 1–17, 2025, DOI: 10.1080/0144929X.2025.2469659.
- [Kh22] Khosrawi-Rad, B. et al.: Conversational Agents in Education – A Systematic Literature Review. In: *ECIS 2022 Research Papers*. Timisoara, Romania, p. 18, 2022, [https://aisel.aisnet.org/ecis2022\\_rp/18](https://aisel.aisnet.org/ecis2022_rp/18).

- [KI24] Klar, M.: How should we teach chatbot interaction to students? A pilot study on perceived affordances and chatbot interaction patterns in an authentic K-12 setting. In: Gesellschaft für Informatik e.V., 2024, DOI: 10.18420/delfi2024.
- [KR20] Kuckartz, U.; Rädiker, S.: Fokussierte Interviewanalyse mit MAXQDA. Schritt Schritt, pp. 55–74, 2020.
- [KS07] Kriz, W. C.; Saam, N. J.: Großgruppenplanspiele als Interventionsmethode. de, Planspiele für die Organisationsentwicklung. Schriftenreihe: Wandel und Kontinuität in Organisationen, 2007.
- [Ku19] Kuckartz, U.: Qualitative Content Analysis: From Kracauer’s Beginnings to Today’s Challenges. en, Forum Qualitative Sozialforschung / Forum: Qualitative Social Research 20 (3), p. 20, 2019, DOI: 10.17169/fqs-20.3.3370.
- [Le24] Lee, U. et al.: Few-shot is enough: exploring ChatGPT prompt engineering method for automatic question generation in english education. en, Education and Information Technologies 29 (9), pp. 11483–11515, 2024, DOI: 10.1007/s10639-023-12249-8.
- [Lu21] Luctkar-Flude, M. et al.: Effectiveness of Debriefing Methods for Virtual Simulation: A Systematic Review. en, Clinical Simulation in Nursing 57, pp. 18–30, 2021, DOI: 10.1016/j.ecns.2021.04.009.
- [Ma24] Malmqvist, L.: Sycophancy in Large Language Models: Causes and Mitigations, 2024, DOI: 10.48550/arXiv.2411.15287.
- [MHC07] Metcalfe, S. E.; Hall, V. P.; Carpenter, A.: Promoting Collaboration in Nursing Education: The Development of a Regional Simulation Laboratory. en, Journal of Professional Nursing 23 (3), pp. 180–183, 2007, DOI: 10.1016/j.profnurs.2007.01.017.
- [Mo22] Molenaar, I.: The concept of hybrid human-AI regulation: Exemplifying how to support young learners’ self-regulated learning. Computers and Education: Artificial Intelligence 3, Number: 100070, 2022, DOI: 10.1016/j.caeai.2022.100070.
- [Na23] Naveed, H. et al.: A Comprehensive Overview of Large Language Models, 2023, arXiv: 2307.06435.
- [Ng23] Ng, D. T. K. et al.: Teachers’ AI digital competencies and twenty-first century skills in the post-pandemic world. en, Educational technology research and development 71 (1), pp. 137–161, 2023, DOI: 10.1007/s11423-023-10203-6.
- [PFS16] Palaganas, J. C.; Fey, M.; Simon, R.: Structured Debriefing in Simulation-Based Education. en, AACN Advanced Critical Care 27 (1), pp. 78–85, 2016, DOI: 10.4037/aacnacc2016328.
- [Re12] Reed, S. J.: Debriefing Experience Scale: Development of a Tool to Evaluate the Student Learning Experience in Debriefing. Clinical Simulation in Nursing 8 (6), e211–e217, 2012, DOI: 10.1016/j.ecns.2011.11.002.
- [Ro57] Rogers, C. R.: The necessary and sufficient conditions of therapeutic personality change. Journal of Consulting Psychology 21 (2), Place: US Publisher: American Psychological Association, pp. 95–103, 1957, DOI: 10.1037/h0045357.
- [Ro62] Rogers, C. R.: The interpersonal relationship: The core of guidance. Harvard Educational Review 32 (4), Place: US Publisher: Harvard Education Publishing Group, pp. 416–429, 1962.
- [SFE16] Sawyer, T.; Flegler, M. B.; Eppich, W. J.: Essentials of Debriefing and Feedback. In (Grant, V. J.; Cheng, A., eds.): Comprehensive Healthcare Simulation: Pediatrics. Springer International Publishing, Cham, pp. 31–42, 2016, DOI: 10.1007/978-3-319-24187-6\_3.

- [Sk24] Skulmowski, A.: Placebo or Assistant? Generative AI Between Externalization and Anthropomorphization. en, *Educational Psychology Review* 36 (2), p. 58, 2024, DOI: 10.1007/s10648-024-09894-x.
- [So25] Song, Y. et al.: Interactions with generative AI chatbots: unveiling dialogic dynamics, students' perceptions, and practical competencies in creative problem-solving. en, *International Journal of Educational Technology in Higher Education* 22 (1), Number: 12, 2025, DOI: 10.1186/s41239-025-00508-2.
- [SQ24] Sommer, B.; von Querfurth, S.: "In the end, the story of climate change was one of hope and redemption": ChatGPT's narrative on global warming. en, *Ambio* 53 (7), pp. 951–959, 2024, DOI: 10.1007/s13280-024-01997-7.
- [SW25] Sun, Y.; Wang, T.: Be Friendly, Not Friends: How LLM Sycophancy Shapes User Trust, 2025, DOI: 10.48550/arXiv.2502.10844.
- [Th11] The Inascl Board Of Directors: Standard VI: The Debriefing Process. en, *Clinical Simulation in Nursing* 7 (4), S16–S17, 2011, DOI: 10.1016/j.ecns.2011.05.010.
- [Th90] Thatcher, D. C.: Promoting learning through games and simulations. *Simulation & Gaming* 21 (3), pp. 262–273, 1990.
- [Ti13] Tilton, K. J.: Non-acute Care Clinical for BSN Students Chronic Illness Care and Diabetes Self-management Support. en, *Doctor of Nursing Practice Systems Change Projects*. 2013.
- [TTH15] Tilton, K. J.; Tiffany, J.; Hoglund, B. A.: Non-Acute-Care Virtual Simulation: Preparing Students to Provide Chronic Illness Care: en, *Nursing Education Perspectives* 36 (6), pp. 394–395, 2015, DOI: 10.5480/14-1532.
- [Ve18] Verkuyl, M. et al.: Comparison of Debriefing Methods after a Virtual Simulation: An Experiment. en, *Clinical Simulation in Nursing* 19, pp. 1–7, 2018, DOI: 10.1016/j.ecns.2018.03.002.
- [VMS21] Verkuyl, M.; MacKenna, V.; St-Amant, O.: Using self-debrief after a virtual Simulation: the process. *Clinical Simulation in Nursing* 57, pp. 48–52, 2021.
- [Wa24] Wang, K.: From ELIZA to ChatGPT: A brief history of chatbots and their evolution. en, *Applied and Computational Engineering* 39, pp. 57–62, 2024, DOI: 10.54254/2755-2721/39/20230579.
- [Za23] Zamfirescu-Pereira, J. et al.: Why Johnny Can't Prompt: How Non-AI Experts Try (and Fail) to Design LLM Prompts. In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. CHI '23, Number: 437, Association for Computing Machinery, New York, NY, USA, pp. 1–21, 2023, DOI: 10.1145/3544548.3581388.