Check for updates

ORIGINAL RESEARCH

The effectiveness and efficiency of using ChatGPT for writing health care simulations

Efrem Violato[1,◉], Carl Corbett[1,2], Brady Rose[3],
Benjamin Rauschning[4], Brian Witschen[5]

[1]Centre for Advanced Medical Simulation, School of Health and Life Sciences, Northern Alberta Institute of Technology, Edmonton, Alberta, Canada
[2]Centre for Advanced Medical Simulation, Northern Alberta Institute of Technology, Edmonton, Alberta, Canada
[3]Medical Laboratory Technology, School of Health and Life Sciences, Northern Alberta Institute of Technology, Edmonton, Alberta, Canada
[4]Paramedicine, School of Health and Life Sciences, Northern Alberta Institute of Technology, Edmonton, Alberta, Canada
[5]Respiratory Therapy, School of Health and Life Sciences, Northern Alberta Institute of Technology, Edmonton, Alberta, Canada

Corresponding author: Efrem Violato, efremv@nait.ca

ABSTRACT

Introduction:
Simulation is a crucial part of health professions education that provides essential experiential learning. Simulation training is also a solution to logistical constraints around clinical placement time and is likely to expand in the future. Large language models, most specifically ChatGPT, are stirring debate about the nature of work, knowledge and human relationships with technology. For simulation, ChatGPT may present a solution to help expand the use of simulation by saving time and costs for simulation development. To understand if ChatGPT can be used to write health care simulations effectively and efficiently, simulations written by a subject matter expert (SME) not using ChatGPT and a non-SME writer using ChatGPT were compared.

Methods:
Simulations generated by each group were submitted to a blinded Expert Review. Simulations were evaluated holistically for preference, overall quality, flaws and time to produce.

Results:
The SME simulations were selected more frequently for implementation and were of higher quality, though the quality for multiple simulations was comparable. Preferences and flaws were identified for each set of simulations. The SME simulations tended to be preferred based on technical accuracy while the structure and flow of the ChatGPT simulations were preferred. Using ChatGPT, it was possible to write simulations substantially faster.

Conclusions:
Health Profession Educators can make use of ChatGPT to write simulations faster and potentially create better simulations. More high-quality simulations produced in a shorter amount of time can lead to time and cost savings while expanding the use of simulation.

What this study adds
- With minimal inputs, ChatGPT can be used to generate outputs that can be used to produce health care simulations efficiently and effectively.
- The outputs are not perfect, and a human subject matter expert is still required to create a final usable simulation.
- Refinement of prompts and practice using ChatGPT can lead to better outputs.
- Utilizing tools like ChatGPT can help reduce workload.

Introduction

Simulation-Based Education (SBE) is essential to Health Professional Education (HPE) [1,2]. Currently, with global needs to provide experiential learning, accelerate time to competency and increase class sizes along with constraints on available clinical/practicum time, the capabilities of simulation make SBE an ever more relevant and important aspect of HPE [3,4]. If SBE is to continue to expand to meet educational and societal needs, there will be a subsequent increase in the demands placed on educators to support SBE, including developing new simulations. Producing and conducting quality simulations can be an intensive and time-consuming process that is done in addition to already heavy workloads [5].

Large Language Models (LLMs) are methods for Natural Language Processing that rely on deep neural networks, specifically the transformer neural network. LLMs process and generate data in sequence to produce text, with the ability to produce novel text from prompts [6]. LLMs and the Generative Artificial Intelligence (GAI) products built off them, particularly ChatGPT (OpenAI, California, USA), are causing excitement, turmoil and discussion across all facets of society with debate from the level of the general public to experts and government, raising questions about application and regulation [7]. For some, there is fear, trepidation and a desire to ban or dismiss the technology, while others are embracing the technology [8]. Much of the discussion has occurred in health care, especially about the performance of ChatGPT on medical examinations [9,10]. This has led to a scramble to regulate and provide guidelines around GAI [7,11]. The expansion of GAI and its influence on HPE are unavoidable [12,13]. Society is currently in the stage of mass adoption of LLM technologies, and amidst the fear and hype, there is the potential for using GAI to improve health care, from education to practice [11,14–16]. As this new technology is here to stay, it is necessary to understand how the potential of GAI can be harnessed for teaching and learning [12,17]. For SBE, this means asking if we can use ChatGPT to improve simulation, a discussion already happening in professional circles [18].

The most immediate application of LLM interfaces like ChatGPT to SBE is in the simulation writing process. ChatGPT may be able to help produce clinical scenarios [16,19] and reduce the time required to write simulations; if effective, one barrier to the expansion of SBE can be removed. To investigate the initial application of ChatGPT to the simulation writing process, two questions were developed:

1. Can ChatGPT be used by a non-subject matter expert (SME) to write health care simulations at an equivalent level to an experienced HPE educator?
2. Can ChatGPT be used by a non-SME to write health care simulations at an equivalent level to an experienced HPE educator in a shorter time?

Methods

Study design

A design approach, specifically feature and usability quality assurance (QA) through Expert Review, was taken to understand if the new method of using ChatGPT leads to a usable, that is, satisfactory and efficient, output [20]. A QA design approach was taken as the development and design of instructional material is a formative and iterative process that best utilizes peer evaluation during alpha development [21]. The present study can be considered an initial use case and exploration of usability and utility or 'prototyping' of a method that is not intended as an absolute or final implementation. To determine the initial utility and usability of ChatGPT to produce simulations for implementation in an HPE program, instructors experienced in simulation compared simulations written by a non-SME using ChatGPT to simulations written by an SME using their standard methods. A non-SME was chosen to write simulations to strengthen the inferences of the study through stress testing, a method to explore the resilience and boundary conditions of function and the limitations of a device or system [22]. If a non-SME can use ChatGPT to write simulations at an equivalent or superior level to a subject matter expert, then it can be inferred that an SME can make even better use of the technology to produce high-quality simulations.

An ethics exemption was provided by the Northern Alberta Institute of Technology Research Ethics Board as the study was determined to fall within the definition of Quality Assurance-Quality Improvement research in accordance with Article 2.5 of the TCPS2.

Simulation development

The Medical Lab Technology (MLT), Respiratory Therapy (RT), and Paramedicine programs were approached to participate in the study as these are simulation-heavy programs. For use in the study, a member of each department responsible for simulation selected a scenario with two to three learning objectives that were being developed by the program. It was requested that scenarios

requiring highly technical skills, for example, pipetting, were not included to allow for a scenario that a non-expert could reasonably develop. The MLT and Paramedicine programs provided one scenario each, while the RT program provided three scenarios.

The scenarios were written by program instructors with a high level of experience designing, writing and conducting simulations. The writer was asked to keep track of the time (hours and minutes) and resources used while writing. The simulation selected for comparison was not indicated to avoid the writer putting a non-normative amount of time or effort into writing a particular simulation. No time limits were set, and writers could work on the simulation until satisfied.

A single scenario writer constructed all the ChatGPT-assisted scenarios. The department representative provided the writer constructing the ChatGPT-assisted scenarios with the same scenario description and learning outcomes as the other writers. The ChatGPT-assisted writer had a background in health professions education and simulation but had no experience in health care practice or specific content knowledge of MLT, RT or Paramedicine. The ChatGPT-assisted writer also tracked the time used to write the scenarios, including time prompting ChatGPT. Within the stress testing framework, the ChatGPT writer was not allowed to access any resources outside of ChatGPT or to use any prompts besides a set of predetermined prompts. To construct the scenario, the ChatGPT-assisted writer prompted ChatGPT and then used the outputs to fill in a simulation template. When producing the simulation, modifications were allowed to be made to the ChatGPT-generated content for coherence and simulation flow; however, no substantive writing could occur. The ChatGPT writer did not view the other scenarios prior to writing the ChatGPT-assisted scenarios.

OpenAI is consistently updating the ChatGPT platform [23] and so for consistency, all simulation generations were done on the same day using the 12 May 2023 version of ChatGPT [24]. The free publicly available version of ChatGPT, running GPT3.5 instead of the subscription version running GPT4, was chosen as anyone can readily utilize the free version.

Materials

Prompts to ChatGPT
Before writing, two members of the research team used an existing simulation to develop prompts to ChatGPT that could provide an adequate amount of information, formatted coherently for creating simulations. The prompts for writing were developed by testing different prompts and comparing the outputs to the information that was contained in the existing simulation. Prompt 3 was developed in part based on the researchers' perspective of a simulation as being conceptualized, written and conducted similarly to the production of a theatrical play, and that ChatGPT can understand topics, language and communication in a similar way to humans.

1. I am going to provide you with a description of a health care scenario, in subsequent communications, this will be referred to as 'The Scenario'. Do you understand?
2. I am now going to provide you with learning objectives, these will be referred to as 'The Learning Objectives'. Do you understand?
3. Based on 'The Scenario' and 'The Learning Objectives' please write me a simulated scenario in the form of a play with scenes, dialogue, and descriptive detail. Please include the patient's vital signs and clinical presentation relevant to 'The Scenario'.
4. Please give me a list of any medical equipment and supplies that would be necessary for conducting this scenario.

Scenarios and learning objectives

All writers used a standardized simulation template that is used for the development of all simulations at the school.

Medical lab technology scenario: Scope

Scenario description: The participant (student) is a new graduate in a small hospital on the night shift. There is one nurse assigned to each ward, and two scheduled for the ER. The participant is called to collect blood work on a patient in the ER who is presumed to be intoxicated. In the room, there is a manikin that is apparently asleep, hooked up to monitoring equipment. As the participant begins to collect, the equipment begins to alarm. A nurse comes in, assesses the patient, and becomes distressed when they realize the patient is now coding. They bring the crash cart over and ask the participant to start an IV while they prepare the other equipment and medication, as the only other nurse scheduled has not yet arrived for their shift, and the physician has been called to another ward to assist in another situation. The nurse becomes irate when the participant explains that they cannot start an IV.

Scenario objectives:

1. Demonstrate awareness of professional boundaries and scope of practice.
2. Maintain composure and respond professionally in ambiguous and emergent situations.

Respiratory therapy scenario: Wards

Scenario description: An RT student has been called to assess an unresponsive patient in the post-surgical ward.

Scenario objectives:

1. Perform a respiratory assessment.
2. Collaborate with the bedside nurse to resolve patient presentation.
3. Provide interventional support.

Respiratory therapy scenario: Transport

Scenario description: Students prepare and transport an ICU patient to CT scan and back.

Scenario objectives:

1. Prepare a patient for transport.
2. Monitor and maintain the patient throughout transport.

Respiratory therapy scenario: Hyper

Scenario description: A patient, in the OR, is starting to show signs of malignant hyperthermia.

Scenario objectives:

1. Recognize the onset of malignant hyperthermia and communicate to the anaesthesia provider.
2. Initiate and/or participate in malignant hyperthermia crisis management.

Paramedicine scenario: Anaphyl

Scenario description: The male or female patient will be presenting with severe upper airway stridor and wheezing. The patient had no known history of allergies or other medical problems prior to today's event. The patient was eating at a local restaurant and a few minutes later, began to present with symptoms.

Scenario objectives:

1. Demonstrate effective crisis resource management/interprofessional collaborative practice skills throughout the simulation.
2. Recognize the symptoms consistent with an anaphylactic reaction.
3. Demonstrate appropriate use of epinephrine and other associated drugs in a severe presentation of anaphylaxis.

Measures

The scenarios and measures for the study were hosted on Qualtrics [25]. Closed and constructed response items were included in the study. Respondents only rated the scenarios relevant to their program. The scenarios were presented identically and blinded with no indication of who wrote them. Raters first completed a set of demographic items to understand their knowledge and experience with simulation. Raters were then asked (1) to select which scenario they would choose to implement and through a constructed response (CR) item to explain their choice; (2) the quality of the simulations rated on a 1–5 Likert Scale from 1 – Very Poor to 5 – Very Good; and (3) to identify any flaws in the scenarios and anything that should be changed, explicated through a CR item.

Participants

Based on the use of Expert Review, three to five respondents were determined to be an adequate sample size for each program group. During formative feature and usability evaluation, an Expert Review with five users will be able to identify ≥80% of errors and issues, a design with larger samples provides diminishing returns and only serves to identify variations of the same issues or phenomena without adding additional insight or value [26,27].

The scenarios and questions were distributed to all program members involved in simulation, except for the writers, through an email sent out by the Program Chair. The SME scenario writers were not involved in any part of the scenario review and evaluation. Participants were informed that the purpose of the QA was to improve how simulations are produced. Completion of the QA was entirely voluntary; no compensation was provided to participants.

Analysis

For the analysis, the lens of design, QA and Expert Review, qualitative methods, were used to understand the quality of the simulations produced, errors and use issues, and how the simulations could be improved [28]. Likert-scale items were treated as ordinal and analysed as counts. All CR items were reviewed directly with no modification to the responses. CR items were addressed using a qualitative descriptive, direct realist approach [29,30] Common themes were identified based on the target and frequency of comments in the CR items.

Results

Time and resources

Substantially less time was required to write simulations using ChatGPT, with a mean difference of 154.8 minutes (2.58 hours) per simulation. Excluding the MLT simulation, which took the human writer 2.45 times longer to write than the other simulations, the mean difference was 112 minutes (1.87 hours). In total, it took 774 minutes (12.9 hours) less to write five ChatGPT-assisted simulations compared to the five human-written simulations (see Outputs, Supplemental Digital Content 1, which contains all prompts and ChatGPT outputs and Scenarios, Supplemental Digital Content 2, which contains all simulation scenarios). The total time to write five ChatGPT-assisted simulations was less than the time for writing the MLT simulation and one of the RT simulations (Table 1). The resources human writers used included discussion and consultation with colleagues, professional competency profiles and professional websites.

Outcomes across programs

Overall, the expert reviewers were highly experienced in simulation (Table 2). Across 13 raters, five MLT experts assessed one MLT scenario, five RT Experts assessed three RT scenarios, and three Paramedicine experts assessed one Paramedicine scenario to produce a total of 23 assessments. For the evaluations of the five different simulations across the three different programs the non-SME ChatGPT-assisted simulations were preferentially selected four times, the SME written simulations were preferentially selected 13 times, and the simulations were considered equivalent 6 times. There was only one simulation where the differential in overall quality rating was >5. Overall, the non-SME

ChatGPT-assisted simulations came close in quality ratings to those produced by the SME (Table 3). It does not appear that there was any pattern in the expert reviewer's rating of the quality of the simulations, or critiques of the simulation, based on experience or knowledge of simulation. The primary flaws across all simulations were centred around three themes: equipment, simulation flow and technical details. Compared to the SME simulations, the ChatGPT-assisted simulations tended to be considered better in simulation flow and worse in technical detail.

Analysis by program

MLT simulation

Sample characteristics
Five MLT instructors completed the study. The MLT sample indicated moderate levels of familiarity with simulation pedagogy, design, and facilitation and a high degree of familiarity with simulation in general (Table 2).

For respondents the mean (SD, Median[Range]), number of simulations written was, 5.6(5.9, 6[0–14]), simulations facilitated 38.8(34.3, 20[10–94]) and years of SBE experience 9.8(3.6, 8[7–16]). The MLT respondents could be considered as *experienced* with simulation.

Choice of simulation1
Two MLT instructors indicated that either simulation would be appropriate to use, two selected the SME simulation and one selected the ChatGPT simulation. For overall quality, the instructors rated the ChatGPT simulation two points lower (15) than the SME simulation (17) (Table 3).

Raters that selected H-Scope preferred that there was more flexibility in the scenario and allowed for more *avenues*, giving the student more time to come to a resolution. The primary flaws for H-Scope were the lack of detail in the equipment list and the scenario was difficult to read and seemed *less objective*. Raters selected CGP-Scope because it was more structured, easier to execute,

Table 1: Time for completion of simulation writing

	Chat generation	Sim writing	ChatGPT total	Human total
MLT – Scope	4 min	30 min	34 min (.57 hr)	360 min (6 hr)
PCP – Anaphyl	4 min	42 min	46 min (.77 hr)	135 min (2.25 hr)
RT – Wards	3 min	29 min	32 min (.53 hr)	180 min (3 hr)
RT – Transport	3 min	29 min	32 min (.53 hr)	120 min (2 hr)
RT – Hyper	2 min	25 min	27 min (.45 hr)	150 min (2.5 hr)
Total	16 min	155 min	171 min (2.85 hr)	945 min (15.75 hr)

Table 2: Frequencies for familiarity with different facets of simulation rated on a 1–5 Likert Scale from not at all familiar to highly familiar

		Familiarity					
		1	2	3	4	5	Total*
Sim General	MLT			1	3	1	20
	RT				5	0	20
	Para			1	1	1	12
	Total			2	9	2	52
Sim Pedagogy	MLT			3	2		17
	RT		2	3			13
	Para			1	2		11
	Total		2	7	4		41
Sim Design	MLT		2	1	2		15
	RT			3	2		17
	Para		1		1	1	11
	Total		4	5	1		43
Sim Facilitation	MLT			2	3		18
	RT			2	1	2	20
	Para				2	1	13
	Total			4	6	3	51

*MLT = 5 respondents, maximum potential total = 25.
RT = 5 respondents, maximum potential total = 25.
Para = 3 respondents, maximum potential total = 15.
Overall total max = 65.

Table 3: Frequencies of scoring for the quality of the simulations from 1 Very Poor to 5 Very Good

		Quality of simulation						
		1	2	3	4	5	Total score*	Differential
MLT	H-Scope		1	1	3		17	2
	CGP-Scope		2	1	2		15	
RT	H-Wards		1	1	3		17	5
	CGP-Wards		3	2			12	
	H-Transport			1	2	2	21	12
	CGP-Transport	3	1		1		9	
	H-Hyper	2			2	1	15	2
	CGP-Hyper	1	1	2	1		13	
Para	H-Anaphyl			1	1	1	12	2
	CGP-Anaphyl			2	1		10	

*MLT = 5 respondents, maximum potential total = 25.
RT = 5 respondents, maximum potential total = 25.
Para = 3 respondents, maximum potential total = 15.
1When simulations are being referenced: H represents the SME human written simulation; CGP represents the non-SME ChatGPT-assisted scenarios. The name indicated aligns with the scenario name in the Methods section, for example, H-Scope indicates SME human written MLT Scope scenario.

and *better written* to achieve the simulation outcomes. For example, the CGP-Scope did not branch, though identified clearly defined endpoints, and had more concise 'scenes' with patient vitals (see Supplemental Digital Content 2). The primary flaw of CGP-Scope was rigidity, with fewer path options for the student to *stand up for their scope of practice*. When either was selected, it was because the scenarios were perceived to be identical.

RT simulations

RT sample characteristics
Five RT instructors completed the study. The RT sample indicated moderate levels of familiarity with simulation pedagogy and a high degree of familiarity with simulation design, facilitation and simulation in general (Table 2). For respondents, the mean number of simulations written was 20.2(11.1, 19[3−30]), simulations facilitated 54.4(32.4, 50[20−100]) and years of SBE experience 7.2(3.56, 7[2−12]). The RT respondents could be considered *highly experienced* with simulation.

Choice of simulations
Wards: Four RT instructors selected the SME simulation, and one selected the ChatGPT simulation. For overall quality, the instructors rated the ChatGPT simulation five points lower (12) than the SME simulation (17) (Table 3).

Raters that selected H-Wards were ambivalent about their choice, with some preference for aspects of the structure of H-Wards and some preference for the structure of CGP-Wards. Both simulations seemed easy, and each had components that would be more applicable at different points in training. H-Wards was seen to require more action from the student and was considered more realistic though the scenario flow was hard to read and lacking in clarity and details, including equipment. H-Wards also had issues with the pharmacology of drugs included and the patient's physiological responses based on initial presentation. CGP-Wards had clearer indications about what to do based on

how students responded, for example, '*IF* the participant recommends providing supplemental oxygen the physician *WILL* agree and ask the participant and nurse to begin providing supplemental oxygen' (see Supplemental Digital Content 2). The equipment listed for CGP-Wards could be more specific. A more detailed initial patient history and presentation would be required for the student to understand the situation and make the appropriate choice of intervention clear.

Transport: Four RT instructors selected the SME simulation, and one selected the ChatGPT simulation. For overall quality, the instructors gave the ChatGPT simulation a substantially lower score (9) than the SME simulation (21) (Table 3).

Raters selected H-Transport because it had more guidance and detail, including background information, which made it easier to follow. However, one respondent also noted that CGP-Transport was easier to follow with more information for the facilitator. Minimal flaws were identified with H-Transport. The CGP-Transport scenario was seen to be vague and lacking in many details, with multiple aspects that were incorrect 'There are so many things wrong with the scenario'. The interventions were considered inappropriate with logical inconsistencies around the interventions 'The scenario says to disconnect from vent and put on supplemental O2. This doesn't make sense'.

Hyper: Two RT instructors indicated either simulation would be appropriate to use, two selected the SME simulation and one selected the ChatGPT simulation. For overall quality, the instructors rated the ChatGPT simulation two points lower (13) than the SME simulation (15) (Table 3).

Both scenarios were seen to be lacking detail, though H-Hyper had more history, ventilator settings, vitals and detail about equipment but didn't seem to flow well, and the patient's vitals did not seem to align with the scenario or the anaesthetist's response. More detail could have been given for the manikin set-up and treatment protocol. CGP-Hyper was seen to be more clearly laid out overall, with more

information about what to do at each stage of the simulation though the equipment list was considered too sparse for MH protocol and more detail for patient history, vital signs and clarification of the student's role was required to set the scene.

Paramedicine simulation

Sample characteristics

Three Paramedicine instructors completed the study. The Paramedicine sample indicated moderate levels of familiarity with simulation pedagogy and design and a high degree of familiarity with facilitation and simulation in general (Table 2). For respondents, the mean number of simulations written was 24.3(11.1, 24[9–40]), simulations facilitated 51.3(42.3, 30[24–100]) and years of SBE experience 8.363(10.2, 4[1–20]). The Paramedicine respondents could be considered *experienced* with simulation.

Choice of simulation

Two Paramedicine instructors indicated that either simulation would be appropriate to use, or one selected the SME simulation. For overall quality, the instructors rated the ChatGPT simulation two points lower (10) than the SME simulation (12) (Table 3).

Raters selected Either as both simulations were seen to have an equal number of strengths and weaknesses. The scenario flow in CGP-Anaphyl was clearer and more organized while H-Anaphyl had better patient and confederate background information. H-Anaphyl lacked scripting for the actors and had an excessive amount of information included, making it difficult to locate pertinent information. CGP-Anaphyl felt *incomplete* and expected actions were described, but the appropriate responses did not emerge until later. One rater preferred H-Anaphyl because it had timelines and expected actions included.

Discussion

The non-SME ChatGPT-assisted simulations were produced substantially faster and while not rated as the same quality overall as the SME versions, achieved quality scores close to the SME version with three of the simulations having a quality differential of two points. It was possible for a non-SME using ChatGPT to write simulations that some educators would rather implement or see as an equivalent choice to an SME-produced simulation.

Except for the CGP-Transport simulation, there were no large differences in the expert rater's evaluation of the two versions of the simulations. Shortcomings and flaws were identified for both the human and ChatGPT simulations, though issues were more frequently identified for the ChatGPT simulations, especially for detail and technical accuracy. Some of the contradictory evaluations, such as a preference for more or less detail or structure, indicate subjective preferences for how a simulation is written influence choice.

In the current design, an extreme approach with the goal of stress testing ChatGPT for writing an entire simulation was taken. For actual application, it is not intended that ChatGPT be used for blind generation and cut and paste

to produce simulations; it is still necessary to have a human SME in the loop. Cooperation between humans and technology will help educators produce the best simulations possible. From a Sociotechnical systems perspective, when there is a human–technology interaction, the first consideration should be to make the technology fit the humans' social, cognitive and physical capabilities [31]. After the human–technology interaction is considered, team and organizational contexts and industry, economic and regulatory contexts are considered [31]. When thinking about using ChatGPT to write simulations, the first consideration is if/how ChatGPT can help educators produce better simulations more efficiently, not if/how ChatGPT will produce simulations alone and what are the higher-level social ramifications.

Insights for writing simulations

ChatGPT can be used for 'inspiration' or as a starting point for producing a simulation; highly detailed and usable outputs can be produced from very sparse descriptions, such as the RT scenarios. Starting with a simple idea, ChatGPT can assist a writer by providing a coherently structured scenario that is generally correct from which the writer can build and refine. The frequently identified issue in the ChatGPT simulations of the lack of specific detail about the equipment, vitals and patient history shows that a human SME is essential. The SME will know what details are required in the scenario, what would be extraneous or wrong, and how to optimize the simulation's scope and difficulty based on the learner's level. Currently, ChatGPT would not perform well if asked to target a specific learner, for example, a second-year RT student.

ChatGPT can produce the initial content for a simulation, helping with the often laborious act of writing itself. The SME can ensure that the content is accurate and appropriate, and that the simulation is properly structured. Human and ChatGPT working together can save the human substantial time and effort.

The consideration of level of detail included is important for improving the prompts that are given to ChatGPT. To obtain a patient history, a prompt could be included that queries ChatGPT about the history of the patient that is described in 'The Scenario'. Prompts can be used to generate better outputs and should be experimented with and refined, aiming for clarity and precision. The outputs will only be as good as the prompts. Additionally, ChatGPT can help write dialogue. Dialogue can be requested and then refined by asking for specific dialogue in the context of the simulation from the 'characters' in the simulation.

Guides can be developed for using ChatGPT. This does not imply guidelines, 'guardrails', or regulation but rather a formalized method to most efficiently use the technology to produce content. Guides can first be developed for how to start effectively using ChatGPT to write simulations before becoming more specific. For example, a series of prompts may be defined that work best to initialize queries for producing Interprofessional simulations and another series that works best for creating technical skills-based simulations.

Limitations

There were two primary limitations: (1) the quality of the prompts used. The prompts used were developed to be simple and to be implemented uniformly across the scenarios. This was done for clarity and consistency. Allowing for further queries and refinement of the prompts would have allowed for improvements in the simulation writing process and would better reflect how people can best utilize ChatGPT. For example, a specific learner level was not targeted with the present prompts; however, by modifying the prompts, multiple cases based on the initial scenario could be generated to target different learner levels or to construct different branches within a single simulation.

(2) The quantity of profession-specific training material for the ChatGPT model. The current ability of ChatGPT to produce professions-specific scenarios may be variable. For example, based on the history of medicine, it can be assumed that more digital content exists for medicine than RT, MLT and Paramedicine. This implication would mean that ChatGPT has been trained on less RT, MLT and Paramedicine content and will have less ability to produce profession-specific content. A query of ChatGPT [24] regarding training content for Medicine vs RT, MLT, and Paramedicine returned the response that 'I don't have access to information about the specific breakdown of training data' while adding, 'I have been trained on a diverse mixture of licensed data, data created by human trainers, and publicly available data from various domains. This extensive training allows me to generate responses and provide information on a wide array of subjects, including both medicine and respiratory therapy'. With this consideration, ChatGPT can likely be used to write simulations for almost any health profession but, presently will be most effective for medicine and nursing.

Implications

ChatGPT is a new way to interface with computers, just like the mouse and Graphical User Interface once were. ChatGPT allows humans to interface with computers using natural language to access the vast body of digitized human knowledge and should be seriously considered by all educators producing health care simulations. The demonstrated time savings of using ChatGPT can reduce workload, allowing educators to focus on other areas. Reducing workload also reduces one factor that can contribute to burnout [5]. There is also the opportunity for cost savings, whether in redirecting salaried employees' time to other areas or obviating the need for external consultants to write simulations. If simulations can be written faster using ChatGPT and for less cost, then there is an obligation to learners, and ultimately to patients, to learn to use the technology. For educators, it is a matter of learning how to critically and judiciously make the best use of ChatGPT to help educate and train learners to be better health care practitioners. Not using the technology at hand would be akin to calculating complex unit changes in dosages in your head rather than using a calculator.

Conclusion

A non-SME used ChatGPT to write simulations for three different health care programs and produced simulations that were occasionally preferred and were of nearly comparable quality to a human SME. The simulations were also produced substantially faster than a human writing a simulation alone. The flaws that arose in using ChatGPT to write scenarios can be ameliorated by including an SME in the loop; humans and machines together can optimize the writing process and likely produce more high-quality simulations faster. ChatGPT is a tool that can make human lives easier and can be utilized to assist humans in growing simulation by improving the quality of experiential learning and expanding capacity in health care systems. Notwithstanding concerns about GAI and its implications, currently, these concerns are largely speculative and tend towards 'hype', innovation can be balanced with ethical considerations, and creating and adapting to technological innovation is an inherent human trait [32].

Declarations

Funding

None declared.

Competing interests

The authors have no disclosures or conflicts of interest, financial or otherwise to declare.

Ethics statement

An ethics exemption was provided by the Northern Alberta Institute of Technology Research Ethics Board for Quality Assurance-Quality Improvement research in accordance with Article 2.5 of the TCPS2.

References

1. Lateef F. Simulation-based learning: just like the real thing. Journal of Emergencies, Trauma, and Shock [Internet]. 2010 [cited 2023 Jun 4];3(4):348–52. Available from: https://journals.lww.com/onlinejets/Fulltext/2010/03040/Simulation_based_learning__Just_like_the_real.9.aspx.
2. Khan K, Pattison T, Sherwood M. Simulation in medical education. Medical Technology [Internet]. 2011 [cited 2023 Jun 4];33(1):1–3. Available from: https://pubmed.ncbi.nlm.nih.gov/21182376/.
3. Hayden JK, Smiley RA, Alexander M, Kardong-Edgren S, Jeffries PR. The NCSBN National Simulation Study: a longitudinal, randomized, controlled study replacing clinical hours with simulation in prelicensure nursing education chief officer. Journal of Nursing Regulation [Internet]. 2014;5(2):S3-S40. Available from: www.journalofnursingregulation.comS3
4. Lee BO, Liang HF, Chu TP, Hung CC. Effects of simulation-based learning on nursing student competences and clinical performance. Nurse Education in Practice. 2019;41:102646. doi:10.1016/j.nepr.2019.102646
5. Quilici AP, Bicudo AM, Gianotto-Oliveira R, Timerman S, Gutierrez F, Abrão KC. Faculty perceptions of simulation programs in healthcare education. International Journal of

Medical Education [Internet]. 2015 [cited 2023 Jun 3];6:166. Available from: /pmc/articles/PMC4662865/

6. Manning CD. Human language understanding & reasoning. Daedalus. 2022;151(2):127–38.

7. Panetta A. Canadian AI pioneer brings plea to U.S. Congress: pass a law now. CBC News [Internet]. 2023 July 26 [cited 2023 Jul 25]; Available from: https://www.cbc.ca/news/world/ai-laws-canada-us-yoshua-bengio-1.6917793

8. Bartz D. U.S. advocacy group asks FTC to stop new OpenAI GPT releases. Reuters. 2023.

9. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digital Health [Internet]. 2023 [cited 2023 Jun 3];2(2):e0000198. Available from: https://journals.plos.org/digitalhealth/article?id=10.1371/journal.pdig.0000198

10. Lubell D. . ChatGPT passed the USMLE. What does it mean for med ed? American Medical Association. 2023. Available from: https://www.ama-assn.org/practice-management/digital/chatgpt-passed-usmle-what-does-it-mean-med-ed

11. Masters K. Ethical use of artificial intelligence in health professions education: AMEE Guide No.158. Medical Technology. 2023;45(6):574–584. doi:10.1080/0142159X.2023.2186203

12. Kasneci E, Sessler K, Küchemann S, Bannert M, Dementieva D, Fischer F, et al. ChatGPT for good? On opportunities and challenges of large language models for education Learning and Individual Differences. 2023;103:102274.

13. van de Ridder JM, Shoja MM, Rajput V. Finding the place of ChatGPT in medical education. Academic Medicine. 2023;98(8):867.

14. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing physician and artificial intelligence Chatbot responses to patient questions posted to a public social media forum. JAMA Internal Medicine. 2023;183(6):589-596. doi:10.1001/jamainternmed.2023.1838

15. Ali SR, Dobbs TD, Hutchings HA, Whitaker IS. Using ChatGPT to write patient clinic letters. Lancet Digital Health. 2023;5(4):e179-e181. doi:10.1016/S2589-7500(23)00048-1.

16. Feng S, Shen Y. ChatGPT and the future of medical education. Academic Medicine. 2023;98(8):867–868.

17. Kurzweil R. The singularity is near. New York, NY: Viking; 2005.

18. Sanko J. ChatGPT and other generative AI Considerations for healthcare simulation. Healthy Simulation. 2023 [cited 2023 Jun 4]. Available from: https://www.healthysimulation.com/50116/chatgpt-generative-ai-healthcare-simulation/

19. Munaf U, Ul-Haque I, Arif T. ChatGPT: a helpful tool for resident physicians? Academic Medicine. 2023;98(8):868–869.

20. Nielsen J. QA & UX. Fremont, CA: Nielsen Norman Group. 2013 [cited 2023 Jul 25]. Available from: https://www.nngroup.com/articles/quality-assurance-ux/

21. Morrison G, Ross S, Kemp J. Designing effective instruction. 6th edition. New York, NY: John Wiley & Sons, Inc. 2010.

22. Chan HA. Accelerated stress testing for both hardware and software. Proceedings of the Annual Reliability and Maintainability Symposium. 2004; 346–351.

23. OpenAI. ChatGPT – release notes [Internet]. 2023 [cited 2023 Jun 4]. Available from: https://help.openai.com/en/articles/6825453-chatgpt-release-notes

24. OpenAI. Chat GPT (May 12th, Version) [Large Language Model]. San Francisco, CA. 2023.

25. Qualtrics XM. Qualtrics. Provo, UT: Qualtrics. 2021.

26. Turner CW, Lewis James R, Nielsen J. Determining usability test sample size. In: Karwowski W, editor. International encyclopedia of ergonomics and human factors [Internet]. 2nd edition. Boca Raton, FL: CRC Press. 2006. p. 3084–3088. Available from: https://www.researchgate.net/publication/242156700

27. Creswell JW, Poth CN. Qualitative inquiry and research design choosing among five approaches. 4th edition. New York, NY: Sage Publishing. 2017.

28. Privitera MB. Overview of a human factors toolbox. In: Privitera MB, editor. Applied human factors in medical device design. Cambridge, MA: Elsevier Academic Press. 2019. p. 19–26.

29. Willig C. What can qualitative psychology contribute to psychological knowledge? Psychological Methods. 2019;24(6):796–804.

30. Sandelowski M. Whatever happened to qualitative description. Research in Nursing and Health. 2000;23:334–340.

31. Hendrick HW. Historical perspective and overview of macroergonomics. In: Carayon P, editor. Handbook of human factors and ergonomics in healthcare and patient safety. 2nd edition. Boca Raton, FL: CRC Group Taylor and Francis Press. 2011. p. 45–64.

32. ChatGPT (2023). [Large Language Model] Conversation with ChatGPT about concerns over generative AI. 2023. Available from: https://chat.openai.com [Accessed 24 May 2023].