Journal of Healthcare Simulation

Check for updates

Journal of Healthcare Simulation

ORIGINAL RESEARCH

Prompt design and comparing large language models for healthcare simulation case scenarios

Sara Maaz[1], Sadek Obeidat[1], Cynthia J. Mosher [1,2], Janice C. Palaganas[2], Nouf Alshowaier[3], Mohammed Almashal[3], Khalid Abu Alsaud[3], Anshul Kumar[2], Areej Al-Wabil[3,4]

[1]*Department of Clinical Skills, College of Medicine, Alfaisal University, Riyadh, Saudi Arabia*
[2]*Department of Health Professions Education, MGH Institute of Health Professions, Boston, Massachusetts, United States of America*
[3]*Department of Software Engineering, College of Engineering, Alfaisal University, Riyadh, Saudi Arabia*
[4]*AI Research Center, Alfaisal University, Riyadh, Saudi Arabia*

Corresponding author: Sara Maaz, SMaaz@alfaisal.edu

https://johs.org.uk/article/doi/10.54531/QZGO9534

ABSTRACT

Background
Large language models (LLMs), such as ChatGPT, Bing and Bard, have shown promise in various applications. Their potential in healthcare simulation scenario design remains minimally explored. With the wide adoption of simulation-based education (SBE), there is an opportunity to leverage these LLMs to streamline simulation scenario creation. This study aims to compare the quality of scenarios generated by LLMs and explore their responses based on different prompting techniques.

Methods
Utilizing a mixed methods exploratory sequential comparative design, we conducted a comparative analysis quantitatively and qualitatively of 90 simulation case scenarios generated among ChatGPT-4, Bing Precise and Bard. Scenarios were generated using two prompting techniques: zero-shot prompting and prompt chaining. The quality of all scenarios was rated using the Simulation Scenario Evaluation Tool.

Results
ChatGPT-4 scored best in both zero-shot and prompt chaining case scenarios, with a mean score of 71.25 and 85.09, respectively, compared to Bard (58.40 and 44.27) and Bing Precise (48.67 and 39.65). Qualitative content analyses were additionally conducted to provide additional insights into the quality of the scenarios.

Conclusions
The findings show marked differences in scenario quality across and between models, underscoring the need for targeted prompt design. This study demonstrates the limitations and potential of LLMs in generating healthcare simulation case scenarios.

Submission Date: 29 February 2024
Accepted Date: 25 September 2024
Published Date: 12 May 2025

Healthcare education is continuously evolving, and the integration of innovative technologies holds great promise in enhancing medical education learning experiences. [1,2]. Healthcare simulation has emerged as a vital tool in medical

education, enabling learners to gain practical skills and clinical decision-making experience in a controlled and low-risk environment.[3–6]. Such simulation typically occurs in a controlled but flexible manner, created using a simulation case scenario crafted to learning objectives, clinical patient progressions, ideal observations, staging and moulage instructions, teaching instructions and materials, as well as necessary equipment and human resources.[7] Simulation case scenarios serve as a communication tool for the facilitation team to ensure high quality and standardization of learning. Simulation case scenarios are constructed using real-life experiences, articles and clinical guidelines, all of which may be provided via search engines and large language models (LLMs).

The convergence of artificial intelligence (AI) and healthcare simulation presents a transformative opportunity in health professions education. In the context of simulation case scenario creation using LLMs, prompt design – or the art of crafting prompts to generate desired outputs in AI models – combines machine learning with creative writing and provides a way of designing prompts for natural language processing (NLP) systems to generate patient case scenarios. Designing and curating a diverse set of scenarios can be a resource-intensive task for educators. The use of LLMs may provide educators with a more extensive array of cases tailored to learners and objectives.

The integration of AI in simulation-based education is a rapidly developing field. While there are multiple concept papers pointing towards the use of ChatGPT for case scenario generation, to the date of this manuscript submission, there has been no published study comparing the feasibility and efficacy of different AI models in generating healthcare simulation case scenarios. In collaboration with engineering faculty and simulationists, this study aims to fill this gap by analysing 90 cases using quantitative and qualitative measures to study the extent to which AI models, Bard, Bing and ChatGPT-4, can assist educators in producing effective and engaging healthcare simulation case scenarios. Leveraging generative GPT, we seek to answer the following research question: For healthcare simulation case scenarios, what are the differences in the quality of cases generated by Bard, Bing and ChatGPT-4? Additionally, we seek to answer the following sub-questions: What are the characteristics of cases between zero-shot and few-shot prompting in Bard, Bing and ChatGPT-4? And what are the future implications of our findings?

Background

Large language models: Bard, Bing and ChatGPT-4

LLMs are a category of AI that undergo extensive pretraining on extensive text datasets gathered from a wide array of sources. This pretraining enables them to grasp complex patterns and connections within the data, which they can then use to predict likely words or phrases in a given context and emulate the way humans process language [8,9]. LLMs use NLP where users enter a prompt into a chatbot, and it generates human-like responses. Bard, Bing and ChatGPT-4 are popular chatbot applications that use different models of LLMs. We chose to explore the capabilities of Bard, Bing and ChatGPT as they are currently the most widely used and accessible chatbots currently in use [10] and they have been pre-trained on a massive corpus of medical diagnostic data, enabling them to offer 'intelligent diagnosis' [11,12].

Google Bard, launched in 2023 and powered by machine learning and natural language models (LMs), is a conversational AI chatbot developed to enable users to get insightful and meaningful responses to their prompts and queries. A major characteristic of Bard is that it is powered by the pathways language model (PaLM 2) and trained on web data [13]. With PaLM 2, Bard has been dominant in generating content as it provides large amounts of information [14]. Bard generates extensive training on large amounts of data. As a result of this, compelling responses can be generated to cater for the needs of users. Notably, Google Bard was rebranded as Gemini in December 2023.

Like Bard, ChatGPT stands as a pioneering force in NLP, particularly in the domain of linguistics. First launched in 2022, ChatGPT was developed in partnership with OpenAI and Microsoft. ChatGPT spans the realms of education, holding the potential to handle complex language tasks, quality writing and medical information.

Microsoft provided OpenAI with funding and technical resources to develop GPT-4, and in return, GPT-4 was used to power Microsoft's 2023 search engine, Bing Chat. Bing, similar to ChatGPT, uses GPT-4 as the foundation of its LM. Where it differs from ChatGPT, however, is in its dataset, features and availability. Bing is free to use, with the sole restriction being the need for a Microsoft account, while ChatGPT-4 requires a subscription. Unlike previous versions of ChatGPT, which relied on data up until 2021, Bing has the ability to pull information from the web in real time (note: the current version of ChatGPT-4, which was not available for use at the time of this study's data collection period, has the capability to search the web in real time). In addition to that, when Bing responds to any given prompt, if the information was directly pulled from a website or an online resource, it will acknowledge those sources by citing them and providing links.

Prompt design

A prompt is the initial input given to an LM to generate a response. This input guides the model to produce the desired output [15]. A prompt has three dimensions: identity, intent and behaviour [16]. Identity informs the chatbot what is being requested, which sets the stage for the chatbot's responses. Intent specifies the area of use. Behavior informs the intended application. Using these three dimensions, the prompts were given in the style of: 'Give me a scenario (identity) to be used (behaviour) for healthcare simulation (intent)'.

Prompt design, a blend of art and science, involves carefully designing and refining the inputs or questions to suit the context. Prompt engineering is the study of these iterations. The goal of prompting is to guide an LLM to provide accurate responses or desired behaviours [17]. Prompt design plays a crucial role in medical applications by enabling more effective and efficient interaction

between healthcare professionals and AI-driven systems [17,18]. In the medical field, carefully designed prompts can facilitate accurate diagnosis, aid in treatment decisions and streamline patient care. The literature reveals various approaches that demonstrate how prompt design can optimize the use of LLMs in educational settings. Bozkurt and Sharma [16] share guidelines for a conversational pedagogy for effective teaching and learning through interaction with LLMs, emphasizing the vital role of well-crafted prompts. Their guidelines include defining the objective, providing context and examples, specifying the desired format or structure, requesting key details to be considered, testing and iterating, and considering safety and ethics. White et al. [19] introduce a prompting catalogue to improve prompt design with ChatGPT. Their research establishes a structured framework for the documentation and implementation of prompts. This framework serves as a systematic approach to deliberating solutions for prompting, pinpoints recurrent prompt patterns as opposed to fixating on individual prompt instances and categorizes these patterns to steer users towards more efficient and successful interactions with LMs. Prompt design is emerging as an important skill for healthcare educators. In order to maximize the potential of generative AI in education, it has become crucial for educators to acquire prompt design as a fundamental skill [17,20].

Simulation scenarios

Healthcare simulation case scenarios generally incorporate many standard fields to meet the various needs of the simulation team: educators (directing the scenario, i.e. objectives, progression of patient status, debriefing points, teaching references), technologists (supporting the technology for the scenario, i.e. staging equipment, programming, moulaging) and standardized persons (acting in the scenario, i.e. scripting, behaviours). The Simulation Scenario Evaluation Tool (SSET) was developed to conduct a structured assessment of the quality of written simulation case scenarios, aiding in the improvement and standardization of simulation-based training scenarios. The SSET is composed of six elements, and each element is composed of items that are graded on a scale of 1–5 (see Supplementary Appendix 1, SSET). The SSET was determined highly reliable (ICC coefficient score of 0.93; $p < 0.001$), and its validity was established via content expert consensus using the modified Delphi Method [21]. The SSET was used in this study to assess the quality of written cases based on the same construct it was originally validated on. For this reason, in SBE case scenario creation, the quality of desired outputs depends on the art of prompt design.

This study represents a collaborative effort between engineering faculty, students and healthcare simulation educators to investigate the potential of LLMs in generating healthcare simulation case scenarios, using both standard and modified prompts and employing zero-shot (i.e. one prompt command) and prompt chaining (i.e. consecutive prompts requesting more information). Additionally, this study evaluates the performance of three state-of-the-art AI models, Bard, Bing and ChatGPT-4 – three different LLMs – in generative healthcare simulation case scenarios, shedding light on the implications and applications of these particular LLMs, as well as the value of prompt design.

Methods

We used a mixed methods exploratory sequential comparative design within and between scenarios generated by LLMs that provide conversational generative AI text-based chatbots (Bard, Bing and ChatGPT-4). This occurred in five steps: (1) generating scenarios; (2) rating the quality of each case scenario using the SSET; (3) analysing the quality between Bard, Bing and ChatGPT-4 from the SSET ratings and analysing the quality within Bard, Bing and ChatGPT-4 between zero-shot prompting and prompt chaining; (4) analysing the descriptive measures of open-ended questions assessing each case scenario; and (5) conducting an expert focus group. We describe each step here.

Step 1. Generation of scenarios

Two main methods of prompting LLMs were considered in this study:

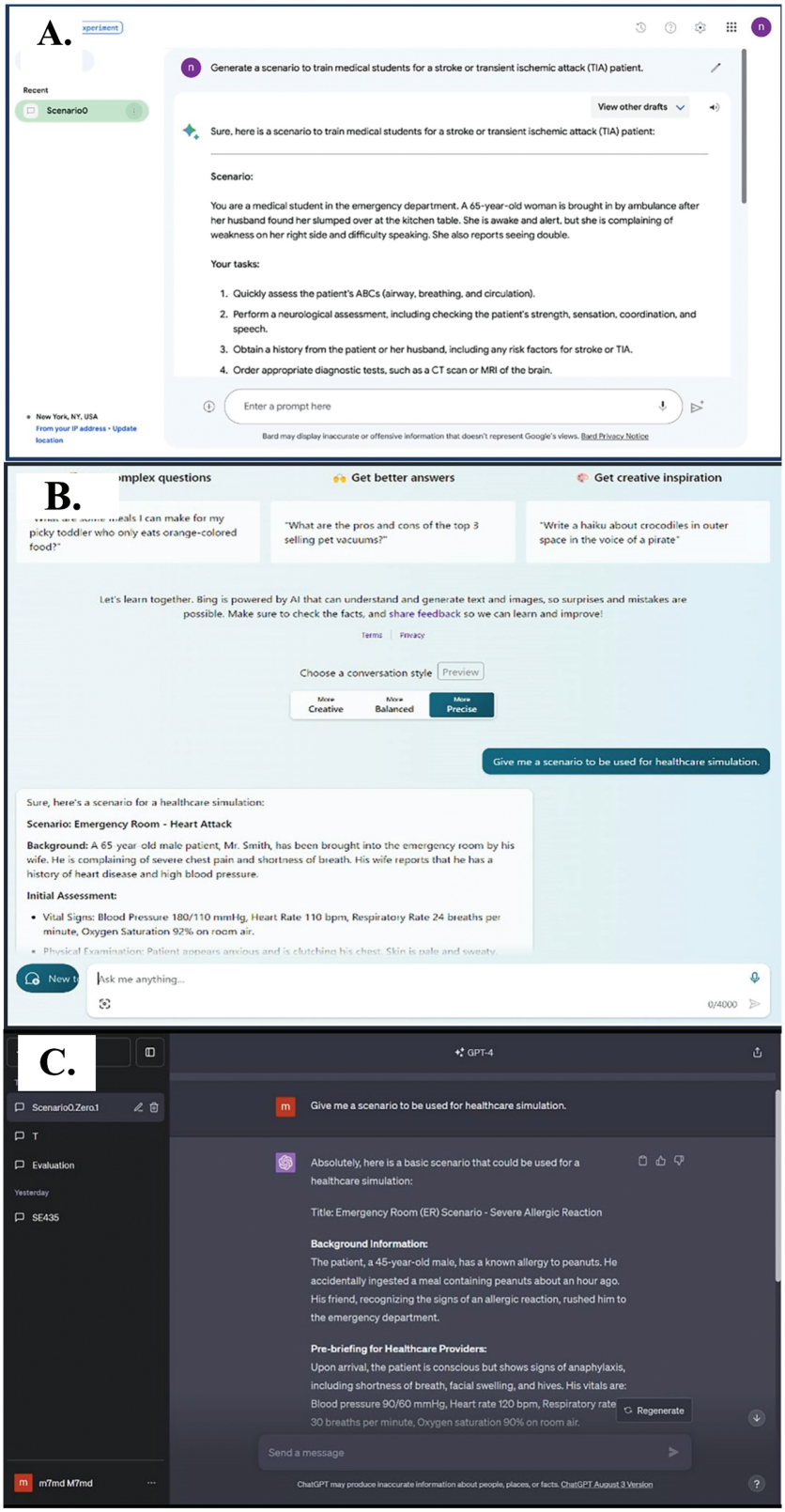
1. Zero-shot: Zero-shot prompting does not require explicit training for a certain task, allowing a model to make predictions about data it has never seen before. In this study, we generated zero-shot simulation case scenarios by giving a single prompt to the selected LMs and not providing examples.
2. Prompt chaining: Prompt chaining uses a number of prompts sequentially with the intent to build from the previous replies. In this study, we generated cases using prompt chaining by starting with the same prompt as the zero-shot cases followed by two additional prompts after the LLM outputs a response.

In our study, we utilized the Google Bard 2.0 version, the April 2023 version of Bing Precise, and the August 2023 version of ChatGPT-4 (see Figure 1).

Data pipeline

To ensure the accuracy and reliability of our study, we established a data pipeline that processed and transformed raw data into an analysis-ready format. The pipeline for prompt discovery comprised several stages, including data preprocessing, data gathering and data categorization. In the preprocessing stage, we studied whether every LLM (ChatGPT-4, Bard or Bing) produced identical responses for identical prompts by repeating the same prompt 10 times while refreshing the conversation in the generative AI chatbot each time. We found that responses are not identical. In fact, responses are different every time. We then compared the responses to ensure consistency in their general content (e.g. semantics, quality, relevance and correctness of model responses based on different prompts). Next, in the categorization stage, we categorized each prompt and response in a separate text-based file that was de-identified for rating. Finally, in the data gathering stage, we copied every prompt with its corresponding response and saved it

Figure 1: (A) Bing, (B) Bard and (C) ChatGPT-4 user interfaces for case scenario generation.



in its designated file. Then, ChatGPT-4 was used to code each case. The coding scheme involved the use of a letter followed by two numbers for each code (e.g. B32). This allowed for the raters to be blinded to which LLM developed the case they were evaluating, removing any potential biases in the evaluation process. The data were then cleaned to eliminate redundant or inconsistent entries and to ensure that all data were in the correct format. Having a precise data pipeline in place not only increased our analytical accuracy but also gave us the ability to extract meaningful insights from our work.

Our procedures for the generation of the scenarios were delineated as follows:

1. We initiated a fresh chat session within each LLM.
2. Subsequently, the specific prompt was presented:

a. **Scenario 0, zero-shot:** 'Give me a scenario to be used for healthcare simulation'.
b. **Scenario 1, zero-shot:** 'Generate a scenario to train medical students for a stroke or transient ischemic attack (TIA) patient'.
c. **Scenario 1 prompt chaining:**
   i. **Prompt 1:** 'Generate a scenario to train medical students for a stroke or TIA patient'.
   ii. **Prompt 2:** 'For that scenario, consider a patient that has a history of an ischemic stroke within the past 3 months'.
   iii. **Prompt 3:** 'For that scenario, add more of the medical history for the patient'.

3. The LLM's response was then captured and saved as raw data.
4. To ensure consistency and prevent any remnants of prior interactions, the following protocol was adopted for clearing conversations: We closed the current web page and launched a new one. After refreshing the page, a new chat session was commenced for the subsequent iteration, following the sequence of zero-shot, and then prompt chaining. This approach prevents the AI from referencing previous interactions within the current session, ensuring that each scenario generation starts afresh. However, it is important to note that this method does not inhibit the AI's overall learning, as updates and learning processes occur on the server side, independent of individual user sessions and IP addresses.
5. For clarity and organization, each chat iteration file was de-identified and labelled by the research assistant for the next step of ratings.

Step 2. Rating of the quality of each case scenario using the simulation scenario evaluation tool

The quality of the AI-generated cases was assessed by four simulation educators using the SSET. The assessors have extensive simulation design backgrounds and have conducted prior studies using the SSET. To ensure reliability, thorough rater training was conducted. We chose two random case scenarios that all raters screened individually, then discussed and came to a consensus as a group, refining our understanding of the SSET. We required three rounds of interrater reliability training to reach a consensus, with a final Cohen's kappa of 0.95. After reasonable reliability was established through consensus, each case was rated by two reviewers independently using the SSET. SM rated all cases to provide consistency in the data. Discrepancies were resolved through weekly meeting discussions.

Step 3. Between- and within-group comparisons: analysing Bard, Bing and ChatGPT-4 across zero-shot and prompt chaining scenarios

Since we had multiple SSET ratings per case, we arranged our data in a long format with two rows of the data spreadsheet for each case (one row per rating for each case). This can be thought of as a repeated measures data structure in which we had two measurements for each case. We then used a linear mixed effects regression model with a random intercept for each case ID to analyse the relationship between the SSET score and the following independent variables: LM, zero-shot prompt case versus prompt chaining case and rater. We also investigated if there are differences in SSET scores for the various combinations of LMs and zero-shot prompt case versus prompt chaining case by including an interaction term for these variables. Our analysis was conducted using R (R Core Team).

Step 4. Qualitative content analysis of case assessments

The research team also conducted two types of qualitative content analysis: (1) open-ended commentary on the assessment of each case regarding suggestions for changes in case writing, and (2) an expert focus group comparing the quality of the prompt chaining cases progressively from zero-shot to third prompt.

At the end of each SSET assessment, an open-ended question was provided to be answered by all raters for all cases: 'What would you change in this case to increase its quality?' We reviewed the entire dataset for familiarization prior to content analysis [22]. We then generated initial labels or codes relevant to each question [23]. The codes were reviewed and categorized continuously and constantly into major categories. Categorized data were checked with each original rater to ensure alignment with rater experience. Recurring responses under each category were noted as themes. We followed this with member checking [21,22] to assess the trustworthiness of the findings by sharing results with the raters to verify their accuracy and alignment with their personal experiences.

Step 5. Qualitative expert focus groups

Four simulationists with a collective 51-year experience (ranging from 5 to 20 years) with simulation case writing individually reviewed each prompt chaining case set (i.e. one case progression from zero-shot, Prompt 2 and Prompt 3 is one 'set'), then met after each case set review to discuss: (a) What are the differences between prompts within each set? (b) What did you find interesting? For analysis, we followed the same content analysis procedure in step 5. We categorized our descriptive observations and thoughts across all focus groups into major categories. Recurring responses under each category were noted as themes. Member checking occurred by sharing quotes with participants to ensure alignment with their thinking.

Results

Quantitative descriptive analysis results

In this part of our analysis, we focused on specific parameters that were chosen for their straightforward analysis and relevance to any written simulation scenario. Our hybrid team of simulation educators and software engineers identified these parameters as they are easily detectable and offer distinguishing features to the cases. By monitoring the number of words, the presence of vital signs, the discussion of symptoms and the presence of identical responses, we established a clear and objective basis for comparing the AI models at a glance.

Scenario 0 – zero-shot prompt

Ten iterations of the prompt, named Scenario 0, zero-shot, 'Give me a scenario to be used for healthcare simulation', were conducted for each LLM.

Bard. Results varied in content and word count. The word count ranged from 246 to 391. Furthermore, 60% of the replies included vital signs, while all of the cases included symptoms and learning objectives.

Bing Precise. In the scenarios analysed, the word count ranged from 139 to 270 words. Each scenario consistently included learning objectives and discussed symptoms. The mention of vital signs was noted in 40% of the scenarios, with half of these mentions being implicit. These implicit references typically involved instructions to measure vital signs as part of the scenario steps, rather than explicitly stating the vital signs in the scenario text.

ChatGPT-4. Word count ranged from 260 to 469. ChatGPT-4 included patient symptoms in all cases. Moreover, learning objectives and vital signs were present in 90% of the generated content.

Scenario 1 – zero-shot prompt

The prompt, 'Generate a scenario to train medical students for a stroke or transient ischemic attack (TIA) patient', was repeated ten times in Scenario 1, zero-shot.

Bard. There were zero identical responses; the content and word count of every result varied. The total number of words varied from 281 to 564. Bard included learning objectives in 80% of its responses. In addition, 30% of the responses included vital signs, and 100% of the cases included patient symptoms.

Bing Precise. Word count had a wide range from 202 to 428. Scenarios provided vital signs 100% of the time. Unlike Bard and ChatGPT, Bing presented us with closely similar responses, as the same patient is mentioned in each scenario but with minor differences in the symptoms and learning objectives. Three cases included links to external resources but did not implement the content into the simulation scenario.

ChatGPT-4. Again, this scenario also generated zero identical responses. The word count ranged from 396 to 572. Learning objectives were mentioned in 7 out of 10 cases. Symptoms were discussed in all the responses (100%). Vital signs were included in 70% of the responses.
There was considerable variability in the cases generated by all three LMs. For more examples, please see Supplementary Appendix 2, Zero-Shot Example Cases.

Scenario 1 – prompt chaining

This experiment was done 10 times with three different consecutive prompts mentioned earlier that were formulated to capture the desired semantic meaning and intent for a healthcare simulation case scenario.

Bard. The word count for the first prompt ranged from 341 to 525. In response to the first prompt, Google Bard offered learning objectives in 60% of its replies, symptom-related information in 100% of its responses and references in 10% of its responses. In addition, 40% of the replies contained vital signs.

The word count for the second prompt varied from 410 to 627. Google Bard responded to the second prompt with learning objectives in 50% of its replies and symptom-related material in all its responses. Furthermore, 40% of the replies contained vital signs.

When we examined the responses, we considered the responses to the three prompts as the complete case. We discovered that the word count ranged from 1226 to 1712. One case began with the setting, time and information about the patient and students. The next prompt provided more information on the scenario itself, and additional patient teaching goals, and then concluded with referenced sources, if any. The second prompt left an impact on only 2 of 10 cases, where the reply had the same structure as the first prompt, but instead of the case being about TIA, the case shifted to recurrent stroke as per our prompt. The patient's instructions, as a result, also changed. The third prompt was helpful in adding a more significant history in only 3 of 10 cases.

Bing Precise. For Scenario 1, across three different prompts, there were noticeable variations in word count and the inclusion of educational components. For Prompt 1, word counts ranged from 250 to 334, with learning objectives present in 60% of responses. All responses included symptoms and mentioned vital signs. In Prompt 2, word counts varied more broadly, from 257 to 409. Here, learning objectives were included in 50% of responses, while symptoms and vital signs were again consistently present in all responses. Finally, for Prompt 3, word counts ranged from 262 to 400. Learning objectives were included in 40% of the responses, but symptoms and vital signs were consistently mentioned in 100% of cases.

Four of the 10 cases considered a diagnosis of recurrent stroke instead of a TIA after the second prompt. Most cases incorporated minimal additional patient history upon the third prompt.

ChatGPT-4. The word count ranged from 418 to 503. Across the three prompts, the scenarios showed consistency in several key educational aspects. Learning objectives and symptoms were included in all responses for the first and second prompts, and in 90% of responses for the third prompt. Vital signs were mentioned in 90% of the responses for the first two prompts and in 80% for the third prompt. The mention of learning objectives varied slightly; in a few instances (specifically in the eighth scenario for each prompt), learning objectives were included during the debrief rather than in the main content.

In 8 of 10 cases, the scenario shifted with the second prompt to focus on a diagnosis of recurrent stroke instead of a TIA. It was also noted that the learning objectives changed in 3 of 10 cases to highlight this shift in focus. After the third prompt, 7 of 10 cases included a more elaborate history.

There was variability in the generated cases for all three LMs. Further examples of cases generated using prompt chaining may be seen in Supplementary Appendix 3.

Figure 2: Comparison of case quality by zero-shot and prompt chaining.

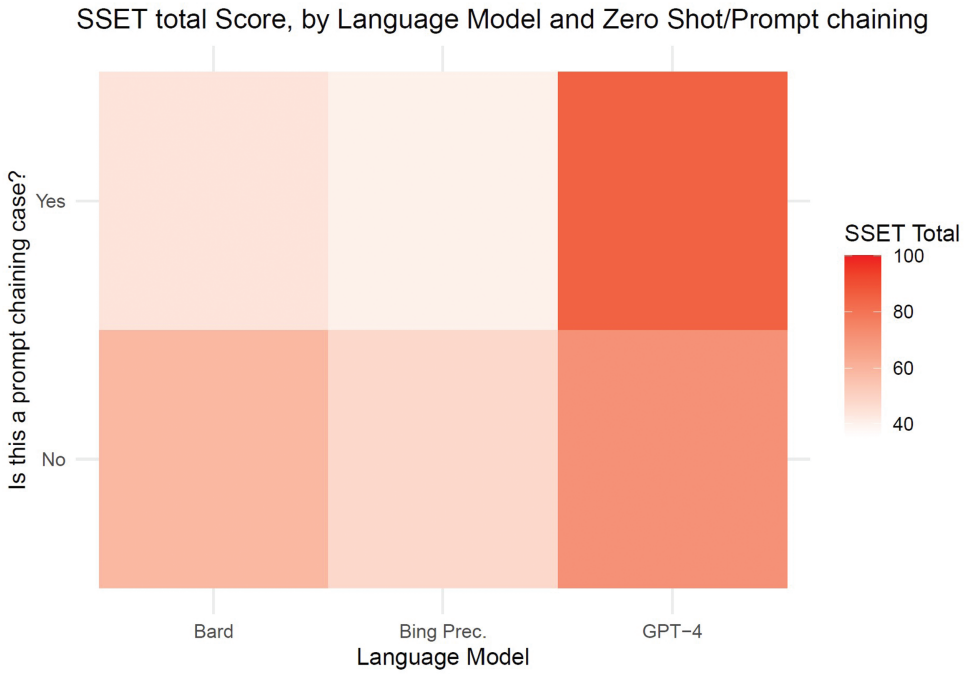


Table 1: Predicted SSET averages from linear mixed effects regression model with interaction

Language model	Is this a prompt chaining case?	Rater 1 SSET average	Rater 2 SSET average	Rater 3 SSET average	Overall average SSET
ChatGPT-4	No	62.14	78.19	73.42	71.25
	Yes	75.98	92.03	87.26	85.09
Bard	No	49.29	65.34	60.57	58.40
	Yes	35.06	51.11	46.34	44.17
Bing Precise	No	39.56	55.61	50.84	48.67
	Yes	30.54	46.59	41.82	39.65

This table contains predicted values from the linear mixed effects regression model in which dummy variables for language model and prompting technique are interacted. For example, the model predicts that, on average, cases in produced by Bard when using prompt chaining were given an average SSET score of 35.06 points by Rater 1.

Comparative analysis results

Figure 2 is a heatmap that shows the overall quality of each of the LLMs for both zero-shot and prompt chaining, based on their SSET scores. The lighter colour on the heatmap represents a lower overall case quality, while the darker colour indicates a better case quality. We see that overall, ChatGPT-4 cases produced the highest quality cases in both zero-shot and prompt chaining categories, with ChatGPT-4 prompt chaining cases being the highest-rated cases across all of the other categories. We also note that Bing Precise produced the lowest quality of cases when compared to the other LLMs. Bard zero-shot and prompt chaining cases sit in between ChatGPT-4 and Bing Precise. Interestingly, the quality of Bard prompt chaining is still lower in quality than ChatGPT-4 zero-shot, which speaks to the difference in the quality of cases produced by ChatGPT-4 and the other LLMs.

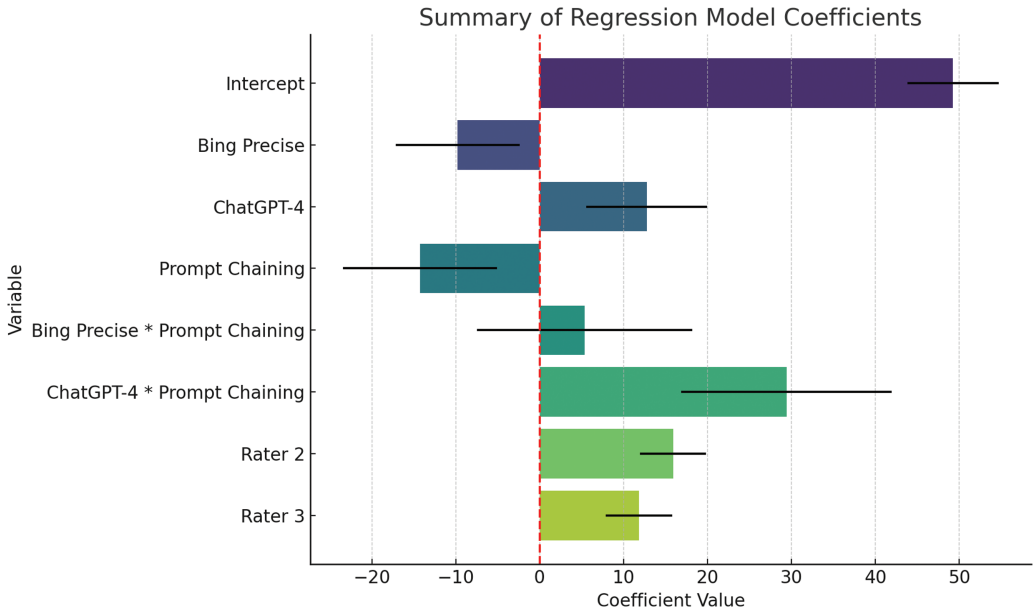
Table 1 presents the average predicted total SSET scores across different combinations of LM, prompting technique (zero-shot/prompt chaining), and rater. These values were calculated using the regression equation from the linear mixed effects regression model with an interaction term between dummy variables for LM and the prompting technique. The full regression equation and the interaction terms are detailed in our data (Supplementary Appendix 4).

Comparing language models

A comparison of the LLMs revealed:

Figure 3 is a visual summary of the regression model coefficients. This bar chart displays the coefficient values for each variable, along with their 95% confidence intervals (indicated by the black error bars). The red dashed line at zero helps to quickly identify which coefficients are positive or negative. Each bar represents the average SSET score for an LM under two different conditions: zero-shot (no prompt chaining) and prompt chaining. The height of each bar reflects the average score achieved by the model, with taller bars indicating better performance.

Figure 3: Summary of regression model coefficients.



- **Bard:** Bard's average score in zero-shot scenarios is 49.29 points, as rated by Rater 1, which serves as a baseline for comparison. However, when prompt chaining is applied, Bard's score drops significantly, indicating a decline in the quality of the generated scenarios. This suggests that Bard may struggle to handle the additional complexity introduced by prompt chaining.
- **Bing Precise:** Bing Precise scores lower than Bard in zero-shot scenarios, reflecting its relatively lower performance. Although the bar for Bing Precise drops further when prompt chaining is applied, the decrease is not as pronounced, and the change is not statistically significant. This suggests that prompt chaining does not have a strong impact on Bing Precise's performance, either positively or negatively.
- **ChatGPT-4:** ChatGPT-4 stands out with the highest scores in both zero-shot and prompt chaining scenarios. The bar for ChatGPT-4 is significantly taller in both conditions, particularly in prompt chaining, where its score increases even further.

These observations underscore the strengths and weaknesses of each model. While Bard and Bing Precise show limitations, particularly under more complex conditions, ChatGPT-4 demonstrates a strong ability to generate high-quality scenarios regardless of the prompt complexity. For those interested in the detailed calculations behind these findings, please refer to the data (Supplementary Appendix 4).

Interaction effects

The interaction between the LM and prompt chaining reveals statistically significant differences in how models perform under varying conditions:

- **ChatGPT-4:** The bar chart shows a substantial increase in ChatGPT-4's performance when prompt chaining is applied. The positive interaction effect here is significant, indicating that ChatGPT-4 not only handles the added complexity well but actually benefits from it. This is evident from the large increase in SSET scores with prompt chaining, making ChatGPT-4 particularly effective in more complex scenarios.
- **Bard and Bing Precise:** Both Bard and Bing Precise exhibit a decrease in performance when prompt chaining is introduced. For Bard, this decline is pronounced, suggesting that the model struggles with the additional complexity of prompt chaining. While Bing Precise also shows a decrease, the interaction effect for Bing is not statistically significant, meaning the change in performance is less conclusive and may not be due to the prompt chaining itself.

To further investigate differences in SSET scores across the three LMs used, we looked separately at the distributions of mean totals for each of the six SSET element groups. These are shown in Figure 4. Even when broken into separate element groups, ChatGPT-4 appears to score the highest on each element.

Also, looking at elements 4 and 5, a drop in scores occurs for all three LMs. Our results suggest that these elements are generally a weak point for AI-written cases, and further investigation into this is needed.

For Bard and Bing Precise, they scored low across all of the elements, with very minimal outliers, indicating their limited use in case writing, whereas prompt formulation in ChatGPT-4 yielded relatively higher results, with several of the cases being almost ready for use to teach students, with a few outlier cases.

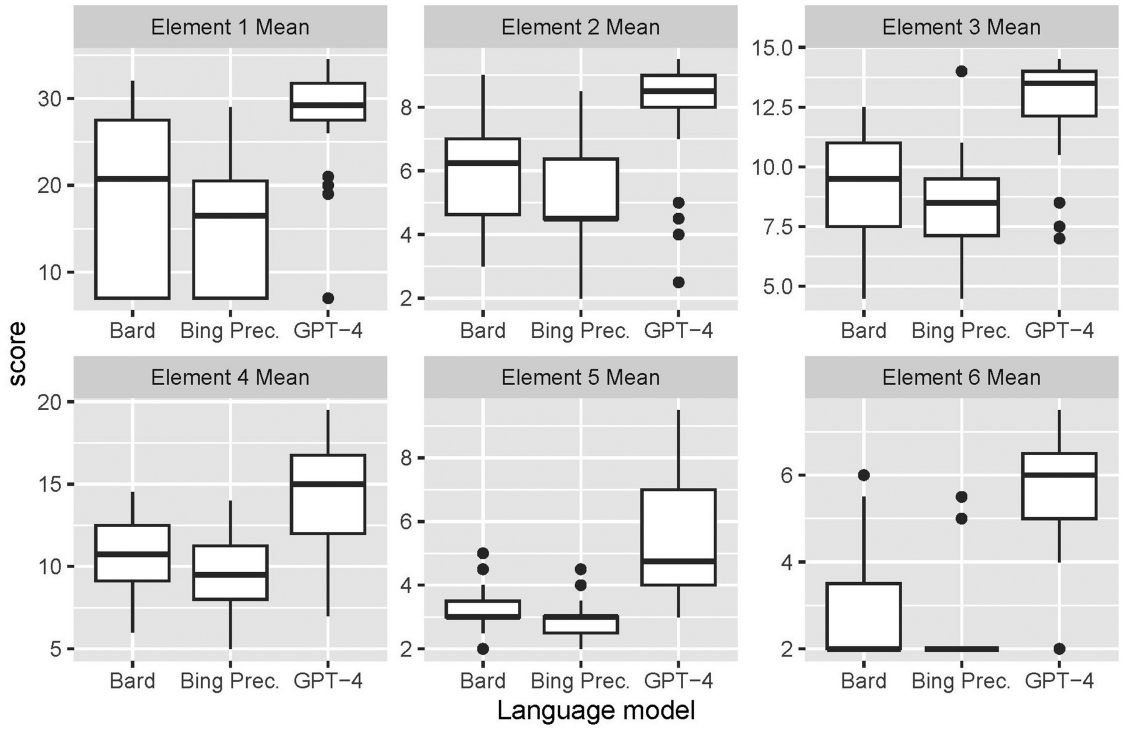
Qualitative content analysis results

Our major categories and subcategories from the analysis of the open-ended question at the end of the SSET assessment or zero-shot cases, 'What would you change in this case to increase its quality?' are listed in Table 2.

Within-group (prompt chaining progression) qualitative content analysis results

A focus group of experienced simulation case scenario writers was conducted. The experts reflected on their

Figure 4: Distribution of individual SSET element scores by language model.



personal experiences with simulation case writing, comparing the LLM-generated cases, which was the intent of this focus group; however, we acknowledge that this brings bias to this section of findings.

Differences in quality

Our data analysis of expert perceptions around the differences between prompts within each prompt chain set revealed two key themes: (1) 'ChatGPT-4 is what I would use', and (2) 'They don't look very different'.

Theme 1: ChatGPT-4 is what I would use

Specific differences were identified anecdotally per each LLM with a consensus that ChatGPT-4-generated cases were superior to those generated by Bard and Bing Precise.

Bard prompt chaining sets. Most cases were seen as 'lacking' in general. In most cases, the addition of a history of stroke didn't reflect on the medical progression of the case nor on the approach to management by students. The third prompt was helpful in adding more significant history, presented as teaching points for students, in three cases only.

Bing Precise prompt chaining sets. All cases were described by the raters as deficient in proper simulation scenario design, necessitating improvements in all elements of the SSET. Most cases incorporated additional minimal patient histories upon the third prompt; however, it was noted that the generated cases were almost identical. The addition of history at the third prompt generally did not contribute significantly to the medical progression or the approach to management by students.

ChatGPT-4 prompt chaining sets. The experts thematically were positive for all ChatGPT-4 cases, stating phrases such as 'one of the best iterated cases', 'good case', 'great case' and 'organized as a sim scenario should be'. Most changes made in cases were deemed 'medically relevant' and 'accurate', and scenario outcomes were changed accordingly with each prompt. The added history with the third prompt was also helpful in most of the cases and adapted to the change in simulation perspective. The group was in consensus that ChatGPT-4-generated cases were closest to human-generated cases.

Theme 2: They don't look very different

The experts felt as though there were minimal differences between prompts. The participant stated, 'It's just one sentence, isn't it? But other than that, everything else is the same'. Every LLM followed the prompt, adding more to each case as directed; however, the experts frequently noted that the additions were not helpful except for the history provided in Bing Precise and Bard upon the third prompt.

Areas of interest

Our data analysis of expert perceptions identified five major themes of interest: (1) 'It's all about the prompting'. (2) 'It looks like a story, not enough to run a sim'. (3) 'The cases are lacking elements', and (4) 'Can I trust the medical accuracy?'

Theme 1: It's all about the prompting

A major theme that emerged in the discussion concerned the art of prompting. Throughout the discussion, the experts consistently suggested that areas found lacking in cases were likely a result of the prompting used. For example, one expert stated, 'The use of the word "scenarios" might have led to more generalized results', giving the cases a story-like appearance. One expert suggested that the use of 'medical students' in the prompt possibly constrained the AI's specification of the target audience.

Table 2: Individual case scenario suggested changes by raters for large language model cases, zero-shot

Category of suggested changes	Subcategory of suggested changes	Response analysis – Bard	Response analysis – Bing Precise	Response analysis – ChatGPT4
Case organizational structure	**Heading and subheading format:** Evaluates the use of appropriate headings and sub-headings for better structure.	16/20 cases needed improvement in their outline. As they lack the main components in simulation design. 2/20 were written as a dialogue.	Most cases were very brief and written in a general descriptive format rather than a simulation case. 18/20 cases needed a better outline.	Most cases used formatting that matches the SSET components. 6 out of 20 needed better case outlines. One case was written as a role play exercise rather than a simulation case. One case was written as a scenario with questions and answers rather than a simulation.
	Target audience Identification: Specifying the level and type of learners is often mentioned.	17/20 cases needed more defined target learners to the simulation.	All cases needed specific target learners.	Although the majority of cases acknowledged the intended learners, a notable proportion, precisely 13/20 cases, necessitates a more detailed specification of the target learners, defining them according to their level of medical education or distinct professional roles such as nurses, residents, or physicians.
Learning objectives	**Specificity:** Clear and specific learning objectives.	12/20 cases needed more specific learning objectives for the simulation. One response specifically mentioned 'well written objectives'.	16/20 cases needed more specific learning objectives for the simulation.	All cases had learning objectives. 7/20 responses highlighted the need for more specific objectives as they were too general for the simulation.
	Appropriateness to the learner level	Since many cases lacked defined target learners, assessing the appropriateness of objectives to learner level was not always applicable. However, one case had objectives that did not match the specific target learner group.	As all cases failed to specify target learners, assessing this aspect was not applicable.	Since many cases lacked defined target learners, assessing the appropriateness of objectives to learner level was not always applicable. However, only 3/20 cases had objectives that did not match the specific target learner group.
Case progression	**Patient states**: How patient states change in branch points in the case.	All cases either completely lacked or had deficient patient states, according to the rater responses.	All cases lacked patient states. Only patient-related details were mentioned in the beginning of the scenario.	Most cases had a set of patient information at the beginning of the simulation and a motion for how the case would resolve after treatment. 18/20 case responses highlighted the need for more elaborate patient states. They lacked detailed patient states at different points during the case.
	Critical actions: measurable actions required at each stage to move to the next stage or to achieve certain outcomes.	17/20 cases needed a critical action list. Some cases had a list of tasks instead of critical actions. It was noted by one of the raters that in some cases, the critical actions didn't seem to have an influence on the flow of the case.	All cases required a clear critical action list to control the simulation's flow. Cases only had some expected tasks without their outcome or related change in patient states.	Most cases outline a list of expected actions to diagnose and manage the patient. But some only had a few general expected actions to be performed as part of the simulation (e.g., take a history, perform a physical examination, and manage), but these actions were not linked to the progression of the case or patient states. 8/10 responses highlighted the need for a list of critical actions that are linked to the patient states for case progression.

continued

Table 2. Continued

Category of suggested changes	Subcategory of suggested changes	Response analysis – Bard	Response analysis – Bing Precise	Response analysis – ChatGPT4
Materials and resources	**Scenario materials:** Comprehensive lists of scenario materials, including equipment, human resources, and operational setup guides.	All responses highlighted the need for a list of materials and resources needed to run the simulation.	All cases lacked a list of resources or any information on what's needed to run the simulation.	All responses highlighted the need for a list of materials and resources needed to run the simulation. In some responses, this was completely lacking, while in others, it just needed elaboration.
	Patient simulation data and audiovisual stimuli	All cases had a brief history of the patient and some key findings from the examination and investigation. But 7/20 were completely lacking in this area	Most cases had scarce histories and findings from patient examinations. Most cases completely lacked lab work and imaging.	Almost all cases provided necessary historical patient data, vitals, physical examination findings, and results of lab work and imaging done. However, in three responses, the cases were lacking in this area.
Debriefing	**Debriefing plan:** Structure a plan with clear objectives.	15/20 cases needed a debriefing plan and objectives.	15/20 cases need a debriefing plan and objectives.	Most cases had debriefing sections. 8/20 responses highlighted the need for an improved debriefing plan.
	Supporting materials	All cases needed supporting materials for debriefing (example: videos or reference materials and clinical guidelines for simulation content).	All cases needed supporting materials for debriefing (example: videos or reference materials and clinical guidelines for simulation content).	All cases needed supporting materials for debriefing (example: videos or reference materials and clinical guidelines for simulation content).
Extra comments	**Use of references or guidelines**	None of the cases used references or guidelines.	One rater identified a case where the AI provided a six-step approach to designing simulation scenario.	Four cases referenced valid medical scales/guidelines. Examples: NIH stroke scale, NRP resuscitation algorithm, PQRST and MONA in myocardial infarction management, and surviving sepsis campaign guidelines.
	Miscellaneous	13/20 cases were described as 'very lacking', 'not written as a simulation case' or 'vague'. 1/20 was described as a 'good case' in terms of design.	Overall weakest generated cases. 16/20 cases were described as 'very lacking' or 'lacking a simulation case structure'.	9/20 cases were described by raters as 'good cases', 'comprehensive' or 'medically accurate' in approach and progression. Only one case was described by a rater as 'very lacking'. One case used a nonexistent English word: 'Fluctify'.

One participant stated, 'In the discussion, they mention if it's a stroke or TIA [but not throughout the case], so it's not consistent! Which is probably a reflection of the prompt itself, because the prompt itself asks for a stroke or TIA, and [the output] is just repeating it'. Another participant stated, 'It takes things literally. There doesn't seem to be a deep understanding of the prompts'. The initial prompt ('Generate a scenario to train medical students for a stroke or transient ischemic attack (TIA) patient'.) was found to be confusing to the participants, where it can be interpreted as *pick either stroke or TIA to train medical students* or *generate a scenario where the medical students need to distinguish whether or not it is a stroke or TIA,* noting, 'Even as humans, we are confused about the prompt'.

Theme 2: It looks like a story, not enough to run a sim

When discussing the differences between prompts within each set, there was a strong sentiment that regardless of LLM or case progression, the cases were 'written like a story'. Unlike human-created simulation cases, the LLM-generated cases were 'missing the branching points that add to the quality of a case'. One expert stated: 'The if and then's [branching points] are what make the cases flexible for student-centered learning'. It was noted that ChatGPT-4, while still missing branching points, provided more structure to cases than Bard and Bing Precise.

In ChatGPT, there's no prediction of what the students are going to do, which is different than say, Bard which

Table 3: Indicators for prompt engineering in LLM-generated case scenarios

Prompt engineering	Examples of indicators
Semantic representation	The brevity of debriefing information suggests that developing prompts that capture the desired semantic meaning and intent for the 'Debriefing Plan' is crucial.
Prompt formulation and structure	The need for understanding the learner's level is critical to ensure that the case is appropriate. This understanding necessitates crafting prompts with appropriate syntax, keywords and context, providing clear instructions to the language learning model (LLM) to specify target learners
Prompt permutations	The observation that the LLMs use a story-telling nature suggests that more exploratory work is needed to generate a diverse set of prompt variations that could help in identifying which phrasings or formulations yield the best results. This could involve systematically modifying sentence structure to include branching points, word order, or incorporating synonyms and paraphrases
Prompt complexity	The lack of content found in the expert focus group suggests more work is needed to explore expanding the prompts with 'prescriptive prose'. Longer prompts may provide more context but may also risk confusing the model, while shorter prompts might lack the necessary context.
Prompt Contextualization	The questionable medical accuracy and the potential for AI hallucinations (false facts) highlight the need for incorporating relevant context or domain-specific information within prompts to enhance the model's knowledge and improve response quality.

> *seemed to have that prediction. You know what I mean? Like, [it says] 'the students are going to do this, and then they're gonna do this.' But you don't really know what the students are gonna do yet.*

While the story-like nature was enjoyable to read, the simulation experts felt that it would not be enough to run a simulation effectively.

Theme 3: The cases are lacking elements

Our data identified lacking elements noted by the experts. This included:

- Information on target learners.
- Clearly labeled and stated objectives, appropriate to the learning.
- References used for the cases, except for diagnostic algorithms.
- Needed materials, equipment and human resources.
- Debriefing plan and resources for debriefing facilitation.
- Human factors considerations.

Part of this may be the disorganization of primarily Bard and Bing Precise:

> *I find it unusable because things are not clearly labeled. Like I'm searching for the objectives within the text.*

> *I really like how ChatGPT has headings. It makes it look more like a case than a story, and I know how to find what I'm looking for.*

The information on target learners was not provided, particularly their experience level or speciality. In many cases, the objectives were not clearly defined but located throughout the case. At times, the objectives were not educational or focused on the target audience but focused on the care of the patient. References were not included in most cases, except for those with recognized diagnostic algorithms. Sections needed for simulating, staging and moulaging were insufficient. The debriefing plan was found to be minimal without suggested structure, scripting, prompts or teaching points.

Other elements noted to be lacking included pre-briefing structure or information for setting the stage, preparation material, information on confidentiality, a guide to establish psychological safety as well as orientation to equipment. These were grouped under human factors considerations.

Scenario, materials and resources

Theme 4: Can I trust the medical accuracy?

An important theme emerged around the medical accuracy of the cases. When changing a diagnosis after the original prompt, the entire case should change appropriately; however, in some sequential iterations, appropriate changes were not made. For example, in this study, a history of a second diagnosis was added to the second prompt in addition to the original prompt diagnosis without fully changing the case. One participant stated, '... the language model wasn't able to diagnose. It did not provide any findings, and so it kept mentioning what was [in the] prompt, and that one should suspect that she is having a stroke or a TIA .... It doesn't mention the distinctive diagnosis'. Another participant associated the vagueness of a case scenario with a lack of medical accuracy: 'It just feels inaccurate because it is so vague. Like they just added the word to add the word since the prompt told it to'.

Discussion

From a proof-of-concept standpoint, our research findings underscore the potential of LLMs in the creation of healthcare simulation case scenarios. It is imperative to emphasize that the quality of the generated output is intricately linked to the specificity and structure of the input. This input critically depends on both the training of simulationists and the science of prompt discovery and prompt design.

Training simulationists is imperative. This involves providing education on LLMs, their functional attributes, capabilities, advantages, disadvantages, limitations and the strategies for using different models. The acquisition of this knowledge is instrumental in recognizing the significance of precise and contextually precise prompting. Additionally,

it is essential to realize that the art of formulating effective prompts, particularly within health professions education, warrants specialized knowledge, skills and considerations. A powerful advantage of LLMs is the contextual adaptability of the system, where LLMs are designed to learn and understand patterns of data and relate them to the context of the given prompt. Semantic representation in prompt design is, therefore, an important consideration for users interacting with LLMs. This involves developing prompts that capture the desired semantic meaning and intent for generating healthcare simulation case scenarios. This often includes exploring semantic role labelling, syntactic analysis for branching algorithms and dependency parsing to create prompts that effectively guide the model's reasoning for healthcare simulation. Examples of prompt design indicators are listed in Table 3. Further exploration and collection of indicators will serve as a useful guide for simulationists to effectively use LLMs for case scenario generation.

Regardless of prompting, ChatGPT-4 was favored overall. This indicated the LM's clear advantage in generating case scenarios. The medical accuracy was noted to be higher in ChatGPT-4 than in Bard, which is interesting as Bard and Bing boast current access and corpus training in medical diagnostics [11,12]. The scenarios generated by ChatGPT-4 aligned more with the characteristics of a typical healthcare simulation scenario that might be created by a simulation educator, surpassing those provided by Bard and Bing Precise.

Overall, the quantitative SSET results revealed high-quality results, while a major theme in the expert focus group was that all cases were seen as 'lacking' and insufficient to use for healthcare simulation. This offers two inferences: there may have been a technology-forgiveness bias, or the SSET does not accurately measure the quality of case scenarios. We believe that both were at play during this study. Because we did not include human-generated cases in our dataset, our best standard may have been set at the best LLM-generated case reviewed versus the ideal human-generated case, setting a lower standard for higher ratings. Additionally, there may be flaws in the instrument for use in evaluating LLM-generated cases. The median score for Bing Precise ranged from 5 to 17 for Element 1 of the SSET. This shows that there was a wide variation in how individual raters rated Bing Precise, which may indicate a high degree of subjectivity. This may be due to a lack of clarity on what makes a well-written case versus a poorly written case.

Simulation cases serve varying purposes within a simulation team, contingent on the roles and responsibilities of team members. The story-like nature of LLM-generated case scenarios may serve students well; however, for simulation educators and scenario designers, revisions of the prompts (i.e. prompt design) are necessary to meet the expected standards and requirements [7] of a simulation scenario for facilitation and debriefing. For simulation educators, this would include learning objectives, critical actions and a debriefing plan. For simulation technology specialists, this would include algorithms or branching points, as well as photos or videos of set-up including staging and moulaging materials.

Limitations

Probably, the most significant limitation of our study was our use of the SSET as a measure of the quality of the case scenarios. This is for two reasons: prompting and missing submeasures. Our expert focus groups suggested that elements of the tool need to be explicitly mentioned in the prompt for a fair evaluation of the LLMs' capability in generating cases that fulfil the SSET standards. Additionally, there were recurring themes in the changes suggested by raters to increase the quality of each case. Many of the suggested changes that the research team felt were important to measuring the quality of a simulation case scenario were not captured by the SSET. These include pre-event information for learners; a clear, organized structure of cases; details of patient states at different points; a clear outline of critical actions, including a list per patient state; ideal participant behaviours for each case progression; references to case information and medical accuracy.

In this study, we compared the quality of cases between different LLMs; however, we did not compare LLM-generated cases with human-created cases. This forgiveness bias may have lowered our rating standard as many elements were noted to have been missing in the LLM-generated cases as compared to our previous experiences with human-created cases. We also did not utilize more advanced promoting techniques like retrieval-augmented generation and few-shot prompting, where one can upload human-created cases to train the system in creating a case in the structure, tone and content of human-created cases. LLMs currently have this capability and may have greatly improved our results.

An additional limitation was the use of standard modes of the LLMs, which may have limited the potential of each. For example, Bing has the option of being creative, balanced and precise. We used the precise mode. ChatGPT, however, is known to have a more creative flare than Bing, likely due to its creative nature in its standard mode. Furthermore, ChatGPT-4 has the option to train the system by uploading your own data. If we had uploaded human-created cases, the output may have been closer to equal quality of human-created cases.

Future implications

Our findings have a number of implications for future research. The most substantial implication is the critical need to understand prompting science to best generate quality case scenarios. Studying the science of prompting may inform educational methods that could best develop simulationists' skills in prompting and ways to create cases using LLMs. For example, studying the various prompt structures that could help inform prompt creation to yield the desired case scenario output may provide practical application tools for simulationists. Specifically, four prompt discovery activities need to be considered in the applied context of healthcare simulation.

1. Prompt effectiveness evaluation: Developing methodologies to quantitatively and qualitatively assess the effectiveness of different prompts in eliciting accurate and relevant responses for healthcare simulation case scenarios.
2. Prompt generalization: Investigating how well a well-optimized prompt for healthcare scenarios can generalize across different LLMs, architectures and datasets.
3. Iterative prompt refinement through active learning: Algorithms can be developed to iteratively learn and refine prompts based on model performance, aiming to reduce human intervention in the prompt design process.
4. Prompt diversity exploration: Analysing the impact of diverse prompts on model behaviour, uncovering potential biases and ensuring fairness in responses.

Another critical implication is the need for humans to carefully review any LLM-generated case, including the medical accuracy. It would be interesting to understand how much human resource time review and verification of cases would require.

Further study is required to determine the superiority of one LLM model when compared to other models. For example, prior to such comparative studies, researchers must understand the differences in the original purpose for each model and the advantages of the use of each as it relates to the study aims, intentionally selecting models that match aims.

This study was conducted in English and did not consider the accessibility of models. It is not known if spoken language generates different results. Google Bard has a wide offering of multilingual support in over 40 various languages, with free access in over 230 countries as compared to ChatGPT, which has 9 languages and 164 countries [24]. Further study in different languages, access and cultures may generate different results.

To gain a deeper insight into the quality of simulation case scenarios generated within LLMs, further research is essential. This research should encompass a comprehensive analysis of overall cases, comparative investigations contrasting LLM-generated cases with those created by humans, and the identification of any missing sub-elements within the SSET. These sub-elements include pre-event information for learners, a well-structured case organization, detailed descriptions of patient conditions at various stages, a clear delineation of critical actions for each patient state, recommended participant behaviours as cases evolve, references to case information and adherence to medical accuracy standards.

Conclusions

This mixed methods exploratory sequential comparative study analysed 90 healthcare simulation case scenarios generated by Bard, Bing Precise and ChatGPT-4, providing a comprehensive view of the capabilities of the different LLMs. While the use of zero-prompt and prompt training with the current version of ChatGPT to create simulation cases does not meet the same quality of human-written scenarios, we found that ChatGPT-4 excelled in comparison to Bard and Bing Precise. Our findings indicate that the quality of the scenarios generated critically depends on the prompting and training of the model, implicating many innovative areas and research considerations for future research in AI-driven tools that provide prompt suggestions, permutations and optimizations based on user-defined criteria for healthcare simulation and objectives. With advanced prompting techniques and evolving LLM technology, LLM use in creating simulation case scenarios may effectively reduce the time and resources typically required for scenario creation. This study contributes recommendations for a systematic process of identifying and optimizing prompts for generating healthcare simulation case scenarios to elicit desired responses from the LLMs.

Supplementary material

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

Acknowledgements

The authors would like to thank Mohamad Bakir and Jared Kutzin for their thoughts on this paper. The authors would also like to thank Alfaisal University for their support in the publication of this work. As our focus is on LLMs, the authors acknowledge their use of AI technology in conducting the study. Finally, we want to thank Rami Ahmed for inspiring us with his work on this topic.

Declarations

Authors' contributions

None declared.

Funding

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Availability of data and materials

The data supporting the findings of this study are available within the supplementary appendices of this manuscript. Should additional data be needed, the authors are open to sharing it upon request.

Ethics approval and consent to participate

None declared.

Competing interests

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Permission to reuse and copyright

Permission for the reuse of the Simulation Scenario Evaluation Tool (SSET) was granted by the British Medical Journal, Request for Use #00793613, on 30 October 2023.

References

1. Guze PA. Using technology to meet the challenges of medical education. Transactions of the American Clinical and Climatological Association. 2015;126:260–270. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4530721
2. Lomis K, Jeffries P, Palatta A, Sage M, Sheikh J, Sheperis C, Whelan A. Artificial intelligence for health professions educators. NAM Perspectives. 2021. doi: 10.31478/202109a
3. Bradley P, Postlethwaite K. Simulation in clinical learning. Medical Education. 2003 Nov;37(11):1–5. doi: 10.1046/j.1365-2923.37.s1.1.x
4. Nestel D, Jolly B, Watson M, Kelly M. Healthcare simulation education: evidence, theory & practice. West Sussex: John Wiley & Sons. 2018.
5. Ziv A, Ben-David S, Ziv M. Simulation-based medical education: an opportunity to learn from errors. Medical Teacher. 2005 May;27(3):193–199. doi: 10.1080/01421590500126718
6. Okuda Y, Bryson EO, DeMaria S Jr, et al. The utility of simulation in medical education: what is the evidence? Mount Sinai Journal of Medicine. 2009;76(4):330–343. doi: 10.1002/msj.20127
7. INACSL Standards Committee. INACSL standards of best practice: SimulationSM simulation design. Clinical Simulation Nursing. 2016 Dec;12:S5–12. doi: 10.1016/j.ecns.2016.09.005
8. Yu P, Xu H, Hu X, Deng C. Leveraging generative AI and large language models: a comprehensive roadmap for healthcare integration. Healthcare (Basel). 2023;11(20):2776. doi: 10.3390/healthcare11202776
9. Yang J, Jin H, Tang R, et al. Harnessing the power of LLMS in practice: a survey on ChatGPT and beyond. arXiv. 2023 Apr 26. doi: 10.48550/arXiv.2304.13712
10. Clusmann J, Kolbinger FR, Muti HS, et al. The future landscape of large language models in medicine. Communications Medicine. 2023 Oct;3(1):141. doi: 10.1038/s43856-023-00370-1
11. Hirosawa T, Mizuta K, Harada Y, Shimizu T. Comparative evaluation of diagnostic accuracy between Google Bard and physicians. American Journal of Medicine. 2023 Nov;136(11):1119–1123. doi: 10.1016/j.amjmed.2023.08.003
12. Caruccio L, Cirillo S, Polese G, Solimando G, Sundaramurthy S, Tortora G. Can ChatGPT provide intelligent diagnoses? A comparative study between predictive models and ChatGPT to define a new medical diagnostic bot. Expert Systems with Applications. 2024 Jan;235:121186. doi: 10.1016/j.eswa.2023.121186
13. Chowdhery A, Narang S, Devlin J, et al. Palm: scaling language modeling with pathways. arXiv. 2022 Apr 5. doi: 10.48550/arXiv.2204.02311
14. Agarwal M, Sharma P, Goswami A. Analysing the applicability of ChatGPT, Bard, and Bing to generate reasoning-based multiple-choice questions in medical physiology. Cureus. 2023;15(6):e40977. doi: 10.7759/cureus.40977
15. Amatriain X. Prompt design and engineering: Introduction and advanced methods. arXiv preprint. arXiv:2401.14423. 2024 Jan 24.
16. Bozkurt A, Sharma RC. Generative AI and prompt engineering: the art of whispering to let the genie out of the algorithmic world. Asian Journal of Distance Education. 2023;18(2):i–vii. Available from: https://asianjde.com/ojs/index.php/AsianJDE/article/view/749
17. Heston TF, Khun C. Prompt engineering in medical education. Int Med Educ. 2023 Aug;2(3):198–205. doi: 10.3390/ime2030019
18. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. New England Journal of Medicine. 2023;388(13):1233–1239. doi: 10.1056/NEJMsr2214184
19. White J, Fu Q, Hays S, et al. A prompt pattern catalog to enhance prompt engineering with ChatGPT. arXiv. 2023 Feb 21. doi: 10.48550/arXiv.2302.11382
20. Li R, Kumar A, Chen JH. How chatbots and large language model artificial intelligence systems will reshape modern medicine: Fountain of creativity or Pandora's box? JAMA Internal Medicine. 2023;183(6):596–597. doi: 10.1001/jamainternmed.2023.1835
21. Hernandez J, Frallicciardi A, Nadir NA, Gothard MD, Ahmed RA. Development of a simulation scenario evaluation tool (SSET): Modified Delphi study. BMJ Simulation & Technology Enhanced Learning. 2020;6(6):344–350. doi: 10.1136/bmjstel-2019-000521
22. Schreier M. Qualitative content analysis in practice. London: SAGE Publications. 2012.
23. Saldaña J. Coding techniques for quantitative and mixed data. In: Onwuegbuzie AJ, Johnson RB, editors.The Routledge reviewer's guide to mixed methods analysis. London: Routledge. 2021. p.151–160.
24. SADA. Bard vs. ChatGPT: A comparison of two leading AI chatbots. SADA Blog. 2023 Jul 17 [cited 2025 Apr 23]; Available from: https://www.sada.com/blog/bard-vs-chatgpt/