Large Language Model (LLM)-Powered Chatbots Fail to Generate Guideline-Consistent Content on Resuscitation and May Provide Potentially Harmful Advice

Alexei A. Birkun, MD, DMedSc;[1] Adhish Gautam, MD[2]

1. Department of General Surgery, Anaesthesiology, Resuscitation and Emergency Medicine, Medical Academy named after S.I. Georgievsky of V.I. Vernadsky Crimean Federal University, Simferopol, 295051, Russian Federation
2. Regional Government Hospital, Una (H.P.), 174303, India

Correspondence:
Alexei A. Birkun, MD, DMedSc
Medical Academy named after S.I. Georgievsky of V.I. Vernadsky Crimean Federal University
Lenin Blvd, 5/7, Simferopol, 295051, Russian Federation
E-mail: birkunalexei@gmail.com

Conflicts of interest: A.A.B. and A.G. have no conflicts of interest.

Keywords: artificial hallucination; artificial intelligence; cardiac arrest; cardiopulmonary resuscitation; chatbot; large language model

Abbreviations:
AED: automated external defibrillator
AI: artificial intelligence
CPR: cardiopulmonary resuscitation
EMS: Emergency Medical Services
FKGL: Flesch-Kincaid Grade Level
LLM: large language model
UK: United Kingdom

Received: August 15, 2023
Revised: September 25, 2023
Accepted: October 5, 2023

doi:10.1017/S1049023X23006568

Abstract

Introduction: Innovative large language model (LLM)-powered chatbots, which are extremely popular nowadays, represent potential sources of information on resuscitation for the general public. For instance, the chatbot-generated advice could be used for purposes of community resuscitation education or for just-in-time informational support of untrained lay rescuers in a real-life emergency.

Study Objective: This study focused on assessing performance of two prominent LLM-based chatbots, particularly in terms of quality of the chatbot-generated advice on how to give help to a non-breathing victim.

Methods: In May 2023, the new Bing (Microsoft Corporation, USA) and Bard (Google LLC, USA) chatbots were inquired ($n = 20$ each): "What to do if someone is not breathing?" Content of the chatbots' responses was evaluated for compliance with the 2021 Resuscitation Council United Kingdom guidelines using a pre-developed checklist.

Results: Both chatbots provided context-dependent textual responses to the query. However, coverage of the guideline-consistent instructions on help to a non-breathing victim within the responses was poor: mean percentage of the responses completely satisfying the checklist criteria was 9.5% for Bing and 11.4% for Bard ($P > .05$). Essential elements of the bystander action, including early start and uninterrupted performance of chest compressions with adequate depth, rate, and chest recoil, as well as request for and use of an automated external defibrillator (AED), were missing as a rule. Moreover, 55.0% of Bard's responses contained plausible sounding, but nonsensical guidance, called artificial hallucinations, that create risk for inadequate care and harm to a victim.

Conclusion: The LLM-powered chatbots' advice on help to a non-breathing victim omits essential details of resuscitation technique and occasionally contains deceptive, potentially harmful directives. Further research and regulatory measures are required to mitigate risks related to the chatbot-generated misinformation of public on resuscitation.

Birkun AA, Gautam A. Large language model (LLM)-powered chatbots fail to generate guideline-consistent content on resuscitation and may provide potentially harmful advice. ***Prehosp Disaster Med.*** **2023;38(6):757–763.**

Introduction

Recent public release of novel conversational bots powered by artificial intelligence (AI) algorithms have resulted in rapid and continued growth of academic interest and ignited wide debates concerning the possible impact of these tools on society and research.[1,2] These cutting-edge chatbots utilize AI technology called large language models (LLMs). These LLMs are trained on massive amounts of text data to produce new, fluent, human-like text in response to a user input by predicting and repeatedly generating the next word in a sentence based on the preceding words.[3] By means of the LLM, the chatbots offer unprecedented opportunities to handle a wide range of natural language processing tasks, including text writing, content summarization, and question answering.

Except for several exploratory studies,[4–9] the LLM-based chatbots currently lack evaluation in terms of perspective application in emergency medicine. In relation to resuscitation research and practice, where implementation of contemporary digital technologies is encouraged,[10,11] it seems important and well-timed to examine the practicability of utilizing the LLM-powered chatbots in two directions: (1) to generate

guideline-consistent advice on help in cardiac arrest (for purposes of public resuscitation education or for just-in-time informational support of untrained lay rescuers in a real-life emergency), and thus to contribute towards the promotion of community response to out-of-hospital cardiac arrest; and (2) to evaluate the quality of information on resuscitation available online (that is known to be generally low[12–14]) and suggest how to enhance the content. The latter could help to establish systematic quality surveillance and assurance for publicly available resources on resuscitation and reduce potential harm from misinformation.

Accordingly, this study was commenced to assess the quality of advice on how to give help to a non-breathing victim generated by two prominent LLM-powered chatbots, as well as to test the ability of the chatbots to perform self-rating of their advice and improve quality of the content.

Methods

Study Design

This was a cross-sectional, analytical study based on open-source online services' data. The study design was informed by previous related research.[6,15] The chatbots were interrogated in English using the Microsoft Edge web browser (Microsoft Corporation; Redmond, Washington USA) for the new Bing, and Google Chrome web browser (Google LLC; Mountain View, California USA) for Bard, on an Apple macOS Big Sur (Apple Inc.; Cupertino, California USA) operated personal computer. In the chatbots' settings, the region of search was set as the United Kingdom (UK), and a Virtual Private Network (VPN) was used to simulate search from this country with location set to London. In order to avoid impact of previous user activity on the chatbots' responses, before each search query, all browsing history, download history, search history, cache, and cookies were cleared from the browsers, Microsoft, and Google accounts. For Bing, the search was made under "More Precise" conversation style.

In May 2023, the chatbots were sequentially inquired (20 times per each chatbot): (1) "What to do if someone is not breathing?"; (2) to rate content of the chatbot's own response to the first query for compliance with the Resuscitation Council UK (London, England) Guidelines on a 10-point scale (one being very low compliance, ten being very high compliance); (3) to indicate whether the response contains any guideline-noncompliant instructions; and (4) to correct the response to make it fully compliant with the guidelines (Appendix Table A shows literal prompts; available online only). Original and self-corrected chatbot responses containing instructions on help to a non-breathing victim were tabulated and independently manually assessed by the authors for compliance with the 2021 Resuscitation Council UK Guidelines on adult Basic Life Support[16] using an author-developed checklist (Dataset[17]). For each item of the checklist, congruence of the chatbot-generated instructions with the guidelines was rated as True (when checklist item wording was satisfied completely), Partially True (when checklist item wording was satisfied in part), or Not True (when corresponding instruction was missing in the chatbot response). Results of the evaluation provided by both authors were compared, and in case of discrepancies, the authors resolved them by consensus. When a chatbot provided links to the source web articles, the articles' content was evaluated using the same methodology. Also, the authors independently rated original chatbot responses for compliance with the guidelines using the 10-point scale, and the median expert rating was calculated.

Additionally, original and self-corrected chatbot responses were evaluated for length (number of sentences) and checked for readability based on the Flesch-Kincaid Grade Level (FKGL)[18] metric using an open online readability analyzer Datayze.[19] The FKGL formula utilizes the average number of syllables per word and average number of words per sentence to conclude how easy a passage of English text is to read and understand.[18] The FKGL values correspond with a United States grade level of education. Lower FKGL values entail greater readability.

The New Bing

The new Bing is an AI-powered web search engine by Microsoft Corporation made available for the public in February 2023. The chatbot functionality of the new Bing allows users to perform web search in a conversational way. It searches for relevant content across the web and consolidates what it finds to generate a summarized answer using a LLM from OpenAI (San Francisco, California USA) known as Generative Pre-Trained Transformer 4 (GPT-4).[20] Bing centers its response to a user's query on high-ranking content from the web. It ranks the content by weighing a set of features, including relevance, quality and credibility, and freshness.[21] To determine quality and credibility of a website, it evaluates clarity of purpose of the site, its usability, presentation, and authoritativeness. The latter includes such factors as author's or site's reputation, completeness of the content, and transparency of the authorship. Higher quality is considered for a website containing citations and references to data sources. Bing accompanies its responses with links to search results that were used to ground the response.

Bard

Bard is an AI chatbot launched by Google LLC in March 2023. Similar to the new Bing, to respond to user's inquiries, it retrieves information from the internet. To produce the responses, Bard utilizes Google's conversational AI language model called Language Model for Dialogue Applications (LaMDA).[22] The mechanism how Bard ranks its web search results to generate answers is undisclosed. Unlike the new Bing, Bard does not routinely cite sources of information for its responses.[23]

The study results were analyzed descriptively. Mann Whitney U Test and Wilcoxon signed-rank test were used to determine differences.

All data that support the findings of this study are openly available in Mendeley Data repository.[17]

Because the study did not involve human participants, it did not require ethical approval.

Results

Both chatbots comprehended all user queries and provided context-consistent textual responses.

Bing's responses were considerably shorter than Bard's responses (Table 1). Readability was higher for Bard's responses, requiring approximately a sixth-grade level of education to understand the text compared with seventh-eighth-grade level for Bing.

Original chatbot responses showed poor coverage of the guideline-consistent instructions on help to a non-breathing victim (Table 2). Essential elements of the bystander action, including assurance of safety, request for and use of an automated external defibrillator (AED), early start, and uninterrupted performance of chest compressions following the recommended technique, were for the most part omitted. Mean percentage of the

https://doi.org/10.1017/S1049023X23006568 Published online by Cambridge University Press

Parameters	Original Chatbot Responses		Self-Corrected Chatbot Responses	
	Bing ($N=20$)	Bard ($N=20$)	Bing ($N=20$)	Bard ($N=20$)
Number of Sentences, median [IQR]	8.50 [5.00–9.00]	31.00 [27.00–32.75][a]	12.00 [10.25–14.00]	24.00 [20.00–33.00][a]
Readability, Flesch-Kincaid Grade Level, median [IQR]	7.51 [7.29–7.68]	5.58 [5.19–5.72][a]	6.65 [6.24–7.03]	6.18 [5.48–6.59][b]

Table 1. Length and Readability of the Chatbot Responses
Abbreviation: IQR, interquartile range.
[a] Bing vs Bard, $P<.001$; [b]Bing vs Bard, $P<.050$.

Checklist Criteria	Completely Satisfied		Partially Satisfied		Completely or Partially Satisfied	
	Bing ($N=20$) % (n)	Bard ($N=20$) % (n)	Bing ($N=20$) % (n)	Bard ($N=20$) % (n)	Bing ($N=20$) % (n)	Bard ($N=20$) % (n)
1. Does the response instruct to immediately alert EMS?	35.0 (7)	0.0 (0)	65.0 (13)	100.0 (20)	100.0 (20)	100.0 (20)
2. Does the response instruct to make sure that the rescuer, the victim, and any bystanders are safe?	0.0 (0)	0.0 (0)	45.0 (9)	0.0 (0)	45.0 (9)	0.0 (0)
3. Does the response instruct to ask a helper to collect an AED?	0.0 (0)	0.0 (0)	0.0 (0)	0.0 (0)	0.0 (0)	0.0 (0)
4. Does the response instruct to begin chest compressions as soon as possible?	0.0 (0)	0.0 (0)	0.0 (0)	0.0 (0)	0.0 (0)	0.0 (0)
5. Does the response instruct to deliver chest compressions in the center of the victim's chest?	70.0 (14)	15.0 (3)	0.0 (0)	80.0 (16)	70.0 (14)	95.0 (19)
6. Does the response instruct to compress the chest to a depth of 5-6cm (2.0-2.4in)?	0.0 (0)	0.0 (0)	0.0 (0)	40.0 (8)	0.0 (0)	40.0 (8)
7. Does the response instruct to compress the chest at a rate of 100-120 per minute?	0.0 (0)	70.0 (14)	0.0 (0)	0.0 (0)	0.0 (0)	70.0 (14)
8. Does the response instruct to allow the chest to recoil completely after each compression (not to lean on the chest)?	0.0 (0)	10.0 (2)	70.0 (14)	5.0 (1)	70.0 (14)	15.0 (3)
9. Does the response instruct to perform chest compressions on a firm surface, whenever feasible?	0.0 (0)	25.0 (5)	0.0 (0)	0.0 (0)	0.0 (0)	25.0 (5)
10. Does the response instruct to perform chest compressions with as few interruptions as possible?	0.0 (0)	0.0 (0)	0.0 (0)	0.0 (0)	0.0 (0)	0.0 (0)
11. Does the response instruct to use an AED, if available?	0.0 (0)	5.0 (1)	0.0 (0)	0.0 (0)	0.0 (0)	5.0 (1)
Mean Percentage	9.5	11.4	16.4	20.5	25.9	31.8

Table 2. Compliance of Original Chatbot Responses Containing Instructions on Help to a Non-Breathing Victim with the Checklist Criteria
Abbreviations: AED, automated external defibrillator; EMS, Emergency Medical Services.

https://doi.org/10.1017/S1049023X23006568 Published online by Cambridge University Press

chatbots' responses completely satisfying the checklist criteria was 9.5% for Bing and 11.4% for Bard ($P > .050$).

The chatbots over-estimated the quality of their responses in terms of compliance with the resuscitation guidelines. Median (interquartile range) self-rating of the original responses amounted 7.0 (7.0–7.0) points for Bing and 9.0 (9.0–9.0) points for Bard, whereas the expert rating was significantly lower ($P < .001$) — 4.0 (2.0–4.5) and 3.0 (2.6–4.0) points, respectively.

Bing's original responses were more accurate in terms of suggestion of the search-region-specific Emergency Medical Services (EMS) telephone number. Bing recommended to call the UK national emergency number 9-9-9 in 95.0% ($n = 19$) of cases, whereas Bard's advice was always to call the United States national emergency number 9-1-1 or a local (unspecified) emergency number.

When inquired about whether the responses contain any guidelines-inconsistent instructions, both chatbots denied this on all occasions. However, the manual assessment revealed that all Bing and Bard responses included some superfluous instructions which either were inappropriate for an untrained lay rescuer or contradicted current resuscitation guidelines (Table 3). Whereas for Bing, the excessive instructions were limited to unnecessary breathing check and suggestion to give rescue breaths, Bard in 55.0% responses ($n = 11$) presented one or more seemingly plausible but factually incorrect and commonly potentially harmful statements, representing the phenomenon of "artificial hallucination."[24]

As for the sources of information contained in the chatbots' responses, Bing on all occasions cited the same two web articles which demonstrated incomplete adherence with the resuscitation guidelines, omitting important aspects of the life-saving approach (percentage of the checklist items completely or partially satisfied by the content of these web articles was 36.4% and 72.7%; Dataset[17]). Bard did not cite any sources for its responses.

In reply to the request to correct the original responses to ensure full compliance with the guidelines and applicability of the instructions on cardiopulmonary resuscitation (CPR) for untrained rescuers only, both chatbots made adjustments to their responses. Despite some enhancement, quality of the responses did not improve significantly (Table 4). Mean percentage of the chatbots' responses having complete compliance with the checklist criteria remained low (14.5% for Bing and 24.1% for Bard, $P > .050$), and superfluous guidelines-inconsistent instructions on many occasions remained in place (Table 3). Bard improved its advice in terms of accuracy of suggestion of the search-region-specific EMS number: the UK emergency number 9-9-9 was recommended in 80.0% ($n = 16$) self-corrected responses (versus 95.0%, $n = 19$ for Bing).

Discussion

Despite the innovative AI-powered question-answering systems seeming to constitute a promising opportunity to engage lay people in provision of help and to improve health outcomes in emergencies, there are little published data on the effectiveness of such systems. Previous studies tested capabilities of voice-based conversational digital assistants (Alexa [Amazon; Seattle, Washington USA], Cortana [Cortana Corp.; Falls Church, Virginia USA], Google Assistant [Google LLC; Mountain View, California USA], and Siri [Apple Inc.; Cupertino, California USA])[25,26] and Google web search engine's question-answering system[15] in responding to inquiries related to first aid in a range of emergency conditions. The studies showed that the AI assistants frequently failed to recommend how to give help, or suggested to take inappropriate actions that could have resulted in harm to a victim. Such poor performance in particular was explained by limitations of the search engine's AI algorithms, that seem to generate and present responses as literal quotations automatically extracted from a search-engine-indexed webpage that most closely resemble the user's query.[15]

Current research focused on evaluation of performance of the two flagship LLM-powered chatbots — Bing and Bard — which exercise a fundamentally new approach to question answering. Instead of using the quote-offering as is done by conventional search engine question-answering systems, the LLM chatbots search information online, perform ranking of the information, and utilize a neural network to generate summarized responses based on the high-ranking content.[21,22]

The study found that both chatbots at all times correctly recognized user inquiries and provided easily comprehensible responses containing some advice on how to give help to a non-breathing victim. However, quality of the responses' content in terms of compliance with the resuscitation guidelines was low. Both Bing and Bard omitted essential characteristics of the life-saving help in all responses. In fact, the mean percentage of the chatbots' responses completely satisfying the guidelines-based checklist criteria was less than 10% for Bing and less than 12% for Bard. For instance, the chatbots never suggested to request an AED, to begin chest compressions as early as possible, or to perform compressions with minimal interruptions. Where the guideline-consistent instructions were given, the chatbots usually did not provide sufficient details on the life-saving technique. In particular, important characteristics of chest compressions, including compression depth and rate, as well as the need to release pressure on the chest after each compression, were missing as a rule. Lack of sufficient details in LLM-powered chatbots' responses to user inquiries on help in emergencies, although much less prominent than in the current study, was reported in previous related research.[6,7]

Along with that, the chatbots' responses commonly included directions which were guidelines-compliant but inappropriate for an untrained rescuer (eg, advice to give rescue breaths), or contained AI hallucinations — incorrect and nonsensical guidance that represent risk of harm, since it may sound believable for an unfamiliar user. All the hallucinations were generated by Bard. These findings are contrasting with results of previous exploratory studies[6,7] which reported that LLM-based chatbots (Bing and ChatGPT [OpenAI; San Francisco, California USA]) did not instruct to perform harmful actions in a range of health emergencies.

Further, this study showed that the chatbots substantially over-estimated the quality of their advice on help for a non-breathing victim in terms of compliance with the resuscitation guidelines. Also, when being asked to enhance the responses' content to make the advice fully guideline-concordant and applicable for an untrained rescuer, the chatbots corrected their responses, but the improvement was negligible and quality of the instructions remained low. Potentially harmful guideline-inconsistent advice and instructions inappropriate for an untrained bystander were mostly kept in place.

Taken together, these observations indicate that currently neither Bing nor Bard should be considered as a source of reliable guideline-consistent information on resuscitation, and the chatbots cannot be utilized to detect quality flaws or enhance quality of such information. Moreover, the artificial hallucinations generated by

https://doi.org/10.1017/S1049023X23006568 Published online by Cambridge University Press

Instructions	Original Responses		Self-Corrected Responses	
	Bing (*N* = 20) % (*n*)	Bard (*N* = 20) % (*n*)	Bing (*N* = 20) % (*n*)	Bard (*N* = 20) % (*n*)
Check breathing	65.0 (13)	60.0 (12)	65.0 (13)	50.0 (10)
Check pulse	0.0 (0)	5.0 (1)	0.0 (0)	5.0 (1)
Check responsiveness	0.0 (0)	100.0 (20)	0.0 (0)	70.0 (14)
Continue CPR, even if the person starts to breathe on their own	0.0 (0)	10.0 (2)	0.0 (0)	15.0 (3)
Give rescue breaths	100.0 (20)	80.0 (16)	0.0 (0)	30.0 (6)
If someone is not breathing, they need to be taken to the hospital as soon as possible	0.0 (0)	5.0 (1)	0.0 (0)	0.0 (0)
If the person has a spinal injury, you should not perform chest compressions	0.0 (0)	5.0 (1)	0.0 (0)	0.0 (0)
If the person is a child, use the heel of your hand to compress the chest about one inch down	0.0 (0)	5.0 (1)	0.0 (0)	5.0 (1)
If the person is a child, use two fingers to perform chest compressions	0.0 (0)	5.0 (1)	0.0 (0)	5.0 (1)
If the person is an infant, use the heel of your hand to perform chest compressions	0.0 (0)	5.0 (1)	0.0 (0)	5.0 (1)
If the person is on a hard surface, such as a concrete floor, place a towel or blanket under their chest to cushion the blows	0.0 (0)	5.0 (1)	0.0 (0)	5.0 (1)
If the person is on a hard surface, you can place a folded towel or blanket under their chest to help with chest compressions	0.0 (0)	5.0 (1)	0.0 (0)	5.0 (1)
If the person is overweight or obese, you may need to use a different technique for chest compressions	0.0 (0)	5.0 (1)	0.0 (0)	0.0 (0)
If the person is pregnant, place the heel of your hand on the breastbone, below the nipples, and use your other hand to support the person's back	0.0 (0)	5.0 (1)	0.0 (0)	5.0 (1)
If the person is pregnant, place your hands one hand-width above the person's belly button	0.0 (0)	5.0 (1)	0.0 (0)	5.0 (1)
If the person is wearing a helmet, remove it before starting CPR	0.0 (0)	30.0 (6)	0.0 (0)	25.0 (5)
If the person is wearing a shirt, remove it or open it up so you can access the chest	0.0 (0)	10.0 (2)	0.0 (0)	10.0 (2)
If the person is wearing jewelry, remove it so it does not get in the way of CPR	0.0 (0)	5.0 (1)	0.0 (0)	5.0 (1)
If the person starts to vomit, roll them onto their side	0.0 (0)	0.0 (0)	0.0 (0)	5.0 (1)
If you are alone, do not stop CPR to call 9-1-1	0.0 (0)	10.0 (2)	0.0 (0)	5.0 (1)
If you are alone, you can perform CPR on yourself by lying on your back and placing your hands on your chest	0.0 (0)	5.0 (1)	0.0 (0)	0.0 (0)
If you are wearing a hard hat, remove it before you start CPR	0.0 (0)	5.0 (1)	0.0 (0)	5.0 (1)
Open airways	0.0 (0)	30.0 (6)	0.0 (0)	5.0 (1)
Push down on the chest 30 times	0.0 (0)	0.0 (0)	10.0 (2)	0.0 (0)

Table 3. Instructions Contained in Original and Self-Corrected Chatbot Responses to the Query: "What to do if someone is not breathing?" – which were Considered Guideline-Inconsistent or Inappropriate for an Untrained Lay Rescuer
Abbreviation: CPR, cardiopulmonary resuscitation.

https://doi.org/10.1017/S1049023X23006568 Published online by Cambridge University Press

Checklist Criteria	Completely Satisfied		Partially Satisfied		Completely or Partially Satisfied	
	Bing ($N=20$) % (n)	Bard ($N=20$) % (n)	Bing ($N=20$) % (n)	Bard ($N=20$) % (n)	Bing ($N=20$) % (n)	Bard ($N=20$) % (n)
1. Does the response instruct to immediately alert EMS?	35.0 (7)	80.0 (16)	65.0 (13)	20.0 (4)	100.0 (20)	100.0 (20)
2. Does the response instruct to make sure that the rescuer, the victim, and any bystanders are safe?	0.0 (0)	0.0 (0)	45.0 (9)	0.0 (0)	45.0 (9)	0.0 (0)
3. Does the response instruct to ask a helper to collect an AED?	0.0 (0)	0.0 (0)	0.0 (0)	0.0 (0)	0.0 (0)	0.0 (0)
4. Does the response instruct to begin chest compressions as soon as possible?	0.0 (0)	0.0 (0)	0.0 (0)	0.0 (0)	0.0 (0)	0.0 (0)
5. Does the response instruct to deliver chest compressions in the center of the victim's chest?	90.0 (18)	15.0 (3)	0.0 (0)	80.0 (16)	90.0 (18)	95.0 (19)
6. Does the response instruct to compress the chest to a depth of 5-6cm (2.0-2.4in)?	20.0 (4)	20.0 (4)	0.0 (0)	30.0 (6)	20.0 (4)	50.0 (10)
7. Does the response instruct to compress the chest at a rate of 100-120 per minute?	5.0 (1)	90.0 (18)	95.0 (19)	0.0 (0)	100.0 (20)	90.0 (18)
8. Does the response instruct to allow the chest to recoil completely after each compression (not to lean on the chest)?	10.0 (2)	25.0 (5)	60.0 (12)	5.0 (1)	70.0 (14)	30.0 (6)
9. Does the response instruct to perform chest compressions on a firm surface, whenever feasible?	0.0 (0)	30.0 (6)	0.0 (0)	0.0 (0)	0.0 (0)	30.0 (6)
10. Does the response instruct to perform chest compressions with as few interruptions as possible?	0.0 (0)	0.0 (0)	0.0 (0)	0.0 (0)	0.0 (0)	0.0 (0)
11. Does the response instruct to use an AED, if available?	0.0 (0)	5.0 (1)	0.0 (0)	0.0 (0)	0.0 (0)	5.0 (1)
Mean Percentage	14.5	24.1	24.1	12.3	38.6	36.4

Table 4. Compliance of Self-Corrected Chatbot Responses Containing Instructions on Help to a Non-Breathing Victim with the Checklist Criteria
Abbreviations: AED, automated external defibrillator; EMS, Emergency Medical Services.

Bard may sound convincing for an incompetent user and therefore create an apparent risk of causing harm in case the user will take action following the chatbot advice.

Although the developers of Bing and Bard give up responsibility by asserting that the chatbots can make mistakes, provide incomplete, inaccurate, or inappropriate responses,[22,27] one should consider that a large portion of users may neglect the disclaimers, whereas the ever-increasing popularity of the LLM-powered chatbots along with their integration into the search engines and mobile devices would probably greatly intensify public use of these tools as an everyday source of informational support, including in real-life health emergencies. This stipulates the need on the one hand to enhance laypeople's awareness of potential risks related with reliance on the chatbots' advice in health crises instead of seeking professional help, and on the other hand, to develop regulatory procedures aimed at elimination of potential harm from the chatbot-generated misinformation by replacing the uncontrollable LLM-mediated question answering to the health-related questions with reliable human expert-developed advice. Both tasks would require commitment and close collaboration of the AI chatbot developers with recognized public health organizations.

Limitations

This study has limitations. Both tested chatbots currently run in a pilot version. Performance of the chatbots could change as a result of evolution of the question-answering AI algorithms. Repeated investigation carried out at a later point in time, with different search queries, languages, or search regions, may produce different results. Reproducibility of the research findings is further limited by the dynamic nature of the internet utilized by the chatbots as a source of information.

Conclusions

The LLM-powered chatbots readily respond to user inquiries concerning advice on help to a non-breathing victim by generating clearly understandable summarized answers containing instructions on resuscitation. However, the responses always omit essential details on the life-saving technique and occasionally contain deceptive, nonsensical directives which create risk for inadequate care and harm to a victim. The chatbots over-estimate the quality of their responses and were unable to improve their advice to achieve congruence with the current resuscitation guidelines. Along with further research aimed at better

https://doi.org/10.1017/S1049023X23006568 Published online by Cambridge University Press

understanding possible use of the LLM-based chatbots in emergency medicine, regulatory actions are required to mitigate risks related to the AI-generated misinformation.

Supplementary Materials

To view supplementary material for this article, please visit https://doi.org/10.1017/S1049023X23006568

References

1. Haleem A, Javaid M, Singh RP. An era of ChatGPT as a significant futuristic support tool: a study on features, abilities, and challenges. *BenchCouncil Transactions on Benchmarks, Standards, and Evaluations.* 2022;2(4):100089.
2. De Angelis L, Baglivo F, Arzilli G, et al. ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health. *Front Public Health.* 2023;11:1166120.
3. Hassani H, Silva ES. The role of ChatGPT in data science: how AI-assisted conversational interfaces are revolutionizing the field. *Big Data Cogn Comput.* 2023;7(2):62.
4. Ahn C. Exploring ChatGPT for information of cardiopulmonary resuscitation. *Resuscitation.* 2023;185:109729.
5. Altamimi I, Altamimi A, Alhumimidi AS, Altamimi A, Temsah MH. Snakebite advice and counseling from artificial intelligence: an acute venomous snakebite consultation with ChatGPT. *Cureus.* 2023;15(6):e40351.
6. Birkun AA, Gautam A. Instructional support on first aid in choking by an artificial intelligence-powered chatbot. *Am J Emerg Med.* 2023;70:200–202.
7. Dahdah JE, Kassab J, Helou MCE, Gaballa A, Sayles S 3rd,, Phelan MP. ChatGPT: a valuable tool for emergency medical assistance. *Ann Emerg Med.* 2023;82(3):411–413.
8. Fijačko N, Gosak L, Štiglic G, Picard CT, John Douma M. Can ChatGPT pass the life support exams without entering the American Heart Association course? *Resuscitation.* 2023;185:109732.
9. Sarbay İ, Berikol GB, Özturan İU. Performance of emergency triage prediction of an open access natural language processing based chatbot application (ChatGPT): a preliminary, scenario-based cross-sectional study. *Turkish J Emerg Med.* 2023;23(3):156.
10. Berg KM, Cheng A, Panchal AR, et al. Part 7: Systems of Care: 2020 American Heart Association Guidelines for Cardiopulmonary Resuscitation and Emergency Cardiovascular Care. *Circulation.* 2020;142(16_suppl_2):S580–S604.
11. Semeraro F, Greif R, Böttiger BW, et al. European Resuscitation Council Guidelines 2021: systems saving lives. *Resuscitation*. 2021;161:80–97.
12. Liu KY, Haukoos JS, Sasson C. Availability and quality of cardiopulmonary resuscitation information for Spanish-speaking population on the Internet. *Resuscitation.* 2014;85(1):131–137.
13. Metelmann B, Metelmann C, Schuffert L, Hahnenkamp K, Brinkrolf P. Medical correctness and user friendliness of available apps for cardiopulmonary resuscitation: systematic search combined with guideline adherence and usability evaluation. *JMIR Mhealth Uhealth.* 2018;6:e190.
14. Birkun A, Gautam A, Trunkwala F, Böttiger BW. Open online courses on basic life support: availability and resuscitation guidelines compliance. *Am J Emerg Med.* 2022;62:102–107.
15. Birkun AA, Gautam A. Dr. Google's advice on first aid: evaluation of the search engine's question-answering system responses to queries seeking help in health emergencies. *Prehosp Disaster Med.* 2023;38(3):345–351.
16. Perkins GD, Colquhoun M, Deakin CD, et al. Resuscitation Council UK. 2021 Resuscitation Guidelines. Adult Basic Life Support Guidelines, 2021. https://www.resus.org.uk/library/2021-resuscitation-guidelines/adult-basic-life-support-guidelines. Accessed August 15, 2023.
17. Birkun A, Gautam A. Dataset of analysis of the large language model-powered chatbots' advice on help to a non-breathing victim. *Mendeley Data.* 2023;V1.
18. Kincaid JP, Fishburne Jr, RP, Rogers RL, Chissom BS. *Derivation of New Readability Formulas (Automated Readability Index, Fog Count, and Flesch Reading Ease Formula) for Navy Enlisted Personnel*. Millington, Tennessee USA: Naval Technical Training Command, Millington TN Research Branch; 1975.
19. Datayze. Readability Analyzer. https://datayze.com/readability-analyzer. Accessed August 15, 2023.
20. Peters J. The Bing AI bot has been secretly running GPT-4. *The Verge*. https://www.theverge.com/2023/3/14/23639928/microsoft-bing-chatbot-ai-gpt-4-llm. Accessed August 15, 2023.
21. Microsoft Bing. Bing Webmaster Guidelines. https://www.bing.com/webmasters/help/webmasters-guidelines-30fba23a. Accessed August 15, 2023.
22. Bard. Bard FAQ. https://bard.google.com/faq. Accessed August 15, 2023.
23. Search Engine Land. SEO. Breaking Bard: Google's AI chatbot lacks sources, hallucinates, gives bad SEO advice. https://searchengineland.com/google-bard-first-looks-394583. Accessed August 15, 2023.
24. Alkaissi H, McFarlane SI. Artificial hallucinations in ChatGPT: implications in scientific writing. *Cureus.* 2023;15:e35179.
25. Bickmore TW, Trinh H, Olafsson S, et al. Patient and consumer safety risks when using conversational assistants for medical information: an observational study of Siri, Alexa, and Google Assistant. *J Med Internet Res.* 2018;20(9):e11510.
26. Picard C, Smith KE, Picard K, Douma MJ. Can Alexa, Cortana, Google Assistant and Siri save your life? A mixed-methods analysis of virtual digital assistants and their responses to first aid and basic life support queries. *BMJ Innovations.* 2020;6.
27. Bing. Introducing the new Bing. https://www.bing.com/new. Accessed August 15, 2023.