Examining the Health Information Quality and Accuracy of Conversational Agents and Generative AI Models in Response to Prompts Regarding Low Back Pain

Leo Li, Alessandra Cory Marcelo, Curtis Cheuk Him Yu, Aaron Tsz Pan Law, Manuela Ferreira

Submitted to: Journal of Medical Internet Research
on: May 12, 2024

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on it's website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressively prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Preprint JMIR Publications

Examining the Health Information Quality and Accuracy of Conversational Agents and Generative AI Models in Response to Prompts Regarding Low Back Pain

Leo Li[1] BASc; Alessandra Cory Marcelo[1] BASc; Curtis Cheuk Him Yu[2]; Aaron Tsz Pan Law[2]; Manuela Ferreira[1] PhD

[1]Sydney Musculoskeletal Health, The Kolling Institute, School of Health Sciences Faculty of Medicine and Health The University of Sydney St Leonards AU
[2]Department of Rehabilitation Sciences The Hong Kong Polytechnic University Hong Kong HK

Corresponding Author:
Leo Li BASc
Sydney Musculoskeletal Health, The Kolling Institute, School of Health Sciences
Faculty of Medicine and Health
The University of Sydney
Level 10, Kolling Building, Royal North Shore Hospital
St Leonards
AU

Abstract

Background: Low back pain (LBP) is a significant global public health concern with a large burden of disease on the population. With the increasing integration of AI technologies in healthcare, it is essential to evaluate their effectiveness in providing high-quality and accurate information when addressing common LBP concerns.

Objective: The purpose of this research is to examine the health information quality and accuracy of conversational agents (CAs) and generative AI (GAI) models in response to questions about LBP.

Methods: A systematic evaluation was conducted on four commonly used CAs and two GAI models using a piloted script of 25 prompts covering various aspects of LBP, including causes, treatment, ability to exercise and work, and imaging. The responses were compiled and transcribed to assess their quality and accuracy. The quality of the responses was assessed using the JAMA benchmark criteria and the DISCERN tool. The accuracy of the responses was assessed by comparing them to the UK NICE Back Pain and Sciatica guidelines and the Australian Lower Back Pain Clinical Care Standard.

Results: The study revealed significant variation in both information quality and accuracy across different CAs and GAI models. Overall, responses exhibited poor quality but moderate accuracy. Siri demonstrated the best overall performance based on a combination of information quality and accuracy scores, whereas voice-only CAs performed the worst in both measures. GAI models had the highest information accuracy but lower information quality overall.

Conclusions: The findings highlight the necessity for improvements in AI health information delivery to ensure the public received reliable and up-to-date information regarding health issues such as LBP. Clinical Trial: N/A

(JMIR Preprints 12/05/2024:60460)
DOI: https://doi.org/10.2196/preprints.60460

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?
✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.
2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

Preprint JMIR Publications

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Preprint
JMIR Publications

Original Manuscript

Preprint
JMIR Publications

Examining the Health Information Quality and Accuracy of Conversational Agents and Generative AI Models in Response to Prompts Regarding Low Back Pain

Abstract

Background: Low back pain (LBP) is a significant global public health concern with a large burden of disease on the population. With the increasing integration of AI technologies in healthcare, it is essential to evaluate their effectiveness in providing high-quality and accurate information when addressing common LBP concerns.

Objective: The purpose of this research is to examine the health information quality and accuracy of conversational agents (CAs) and generative AI (GAI) models in response to questions about LBP.

Methods: A systematic evaluation was conducted on four common used CAs and two GAI models using a piloted script of 25 prompts covering various aspects of LBP, including causes, treatment, ability to exercise and work, and imaging. The responses were compiled and transcribed to assess their quality and accuracy. The quality of the responses was assessed using the JAMA benchmark criteria and the DISCERN tool. The accuracy of the responses was assessed by comparing them to the UK NICE Back Pain and Sciatica guidelines and the Australian Lower Back Pain Clinical Care Standard.

Results: The study revealed significant variation in both information quality and accuracy across different CAs and GAI models. Overall, responses exhibited poor quality but moderate accuracy. Siri demonstrated the best overall performance based on a combination of information quality and accuracy scores, whereas voice-only CAs performed the worst in both measures. GAI models had the highest information accuracy but lower information quality overall.

Conclusions: The findings highlight the necessity for improvements in AI health information delivery to ensure the public received reliable and up-to-date information regarding health issues such as LBP.

Keywords: low back pain; health literacy; public health; consumer health information; conversational agents; generative AI

Preprint JMIR Publications

Introduction

Low back pain (LBP) is a significant global public health concern [1]. Since 2003, LBP has caused the highest burden of disease according to years lived with disability in Australia [2]. The prevalence of LBP is rising as well and is projected to hit 843M cases worldwide by 2050 [3]. Estimates predict that LBP poses an economic burden of $3.4B on the Australian healthcare system annually [4]. This is because LBP is a multifactorial condition with psychological, social, and biophysical factors, comorbidities, and pain-processing mechanisms all contributing to both pain and associated disability [1]. Recent research has shown that there has been a shift from clinician-led management to self-management strategies, where people play an active role in their own recovery [5]. These self-management strategies include lifestyle adjustments, exercise, and use of reliable information to make informed decisions about treatment and future prevention.

With the rapid advancement in internet technologies, Conversational Agents (CAs), described as voice-activated "chatbots" and accessible through internet-connected technologies, have been widely adopted to search for health related information [6, 7]. Commonly used CAs include Apple's Siri, Amazon's Alexa, Google Assistant and Microsoft Cortana, accessible mainly through smartphones, smart speakers, or computers [8-10]. CAs operate by processing user's voice queries, retrieving information from various sources, and then providing either a spoken response or a visual answer on the screen [11]. They are easily accessible and convenient to use, which has resulted in an increasing number of internet searches being conducted using voice search, including searching for health related information [12].

Recently, academic interest in assessing the quality and accuracy of health related information provided by CAs has been sparked by the increasing prevalence of CAs as tools for health information retrieval [13]. In the context of this study, quality is defined as "the state or ability of data being good enough to fulfil the goals and purposes of the analysis" [14]. Information that is high quality is suitable and reliable for the intended purposes. Meanwhile accuracy is referred to as "the correctness of the output information" [15]. Accurate information is measured as information that is true or correct compared to the expected output. For instance, Miner et al. found that CAs from Google, Apple and Samsung responded inconsistently and incompletely when asked questions about mental health, physical health and interpersonal violence [16]. Kocaballi et al. discovered that commonly available, general-purpose CAs on smartphones and smart-speakers are limited in their ability to advise on safety-critical health prompts and lifestyle prompts [9]. In contrast, other studies have found some CAs on certain devices can be a reliable and accurate source of information when asked questions about health information [6, 12]. For example, Alagha et al. found Siri and Google Assistant to high quality and accurate answers to consumer health questions about vaccine safety and use [12]. Goh et al. found similar results for Siri and Google Assistant, this time in response to questions seeking COVID-19 related information [6]. A recent scoping review on CAs in healthcare found that results from studies evaluating CAs were generally positive, but called for more robust evaluations of these CAs, focusing on their safety and effectiveness in delivering information [10]. Collectively, the scientific literature shows that the quality of health information provided through CAs is highly variable, with some low quality, biased, or even misleading health information provided to the patients [6, 7, 9, 12, 16]. This could lead to high risk of development of exacerbation of a patient's condition if they have followed wrong instructions provided by low-quality information. Nevertheless, little research has been conducted in examining how well CAs perform in relation to queries related to LBP. Due to the impact that low quality and inaccurate information may result in LBP patients, it is vital to examine the health information quality and accuracy of CAs in response to prompts regarding LBP, an area which has been overlooked in the literature. Thus, the first objective of the study is to examine the information quality and accuracy of CAs when responding to prompts regarding LBP.

The second objective of this study is to assess how well Generative AI (GAI) can assist patients in answering questions related to LBP, in comparison with CAs. The motivation for this second research objective is driven by the rapid rise in the popularity of ChatGPT, a type of GAI, since November 2022 [17]. This research study focuses on Generative Pre-Trained Transformer (GPT) GAI models, which are designed to generate detailed natural language responses to a natural language prompt in a conversational way [18, 19]. Two popular GAI models are ChatGPT from OpenAI, and BingChat from Microsoft, both generating content by analysing and learning from non-real time mass data sets [20]. The limited research in this area shows that GAI models such as ChatGPT-3.5 are able to demonstrate highly accurate responses to various medical questions across specialities [21]. Literature further proposes that the utilisation of GAI models in healthcare may be significantly increased in the future, with great potential to provide more precise and personalised treatment recommendations to the patients [21-23]. However, to the best of our knowledge, there have been no studies examining how GAI models respond to prompts regarding LBP, especially comparing to CAs. Therefore, this research study aims to examine whether GAI models can produce more higher quality and more accurate responses than CAs in response to prompts regarding LBP.

This study aims to examine the health information quality and accuracy of CAs and GAI models in response to questions about LBP. In order to achieve the aim, the following three research questions are answered.

1. What is the quality of the information provided by CAs and GAI models in response to prompts about LBP?
2. How accurate is the response of CAs and GAI models to prompts about LBP?
3. How do GAI models perform compared to CAs in regards to quality of responses and accuracy to prompts about LBP?

Methods

Study Design and Sample

This study employed an observational, cross-sectional design to evaluate the quality and accuracy of the health information provided by CAs and GAI models in response to prompts regarding LBP. The performance of four commonly used CAs, including two voice-only devices (i.e., CAs that run on devices without a screen), Amazon’s Alexa on the Amazon Echo Dot 2 and Google Assistant on the Google Home (hereafter referred to as Google Assistant – Home), and two multimodal devices (i.e., CAs that run on devices with a screen), Apple’s Siri on the iPhone and Google Assistant on the Google Pixel (hereafter referred to as Google Assistant – Pixel) were assessed. Multimodal conversational agents were included as they are able to present responses in multiple forms with supplementary information in response to queries [6]. Additionally, two popular GAI models: OpenAI’s ChatGPT-3.5 and BingChat on the Microsoft Edge were assessed in the study. These devices and models were chosen as they are widely accessible and commonly used [24]. This targeted sampling approach was also guided by their relevance to the purpose of this study [25]. (Table 1) summarises the characteristics of the study samples.

Table 1: Study sample characteristics.

Device Name	Type (GAI	Characteristics (Voice Only/Multimodal/Text)	Working Principle	Search Source

Preprint JMIR Publications

	or CA)			
Alexa	CA	Voice Only	Natural Language Processing (NLP) [26]	Bing Search [27]
Google Assistant – Home	CA	Voice Only	NLP [26]	Google Search [28]
Siri	CA	Multimodal	NLP [26]	Google Search [28]
Google Assistant – Pixel	CA	Multimodal	NLP [26]	Google Search [29]
ChatGPT-3.5	GAI	Text Only	Large Language Model (LLM) [30]	Reinforcement Learning from Human Feedback [19]
BingChat	GAI	Text Only	LLM [31]	Bing Search [32]

As the study did not include a human research population, no application was required to be submitted to the ethics committee.

Data Collection

Pilot Study

To ensure the rigour of the study, a pilot study was conducted to ensure that the prompts used in the main study would be correctly understood by the CAs and GAI models and elicit a valid response [9]. The procedure is illustrated in (Figure 1).

[Insert Figure 1]

The first step of the pilot study was to develop the initial set of prompts for the study. These prompts were first derived from a comprehensive literature review on common patient concerns regarding their LBP. These concerns were examined by the research team and organised into four main categories, included “Causes of LBP”, “Management and Treatment of LBP”, the ability to “Work and Exercise with LBP”, and “Imaging for LBP” [33, 34]. A preliminary set of 28 prompts were formulated from the insights of the literature review. To ensure the methodological rigour of the study, the prompts were then validated through consultation with five practicing clinicians with experience in managing patients with musculoskeletal conditions [9]. The clinicians were presented with the prompts and asked to confirm whether they were similar to usual concerns voiced by patients during treatment of LBP. The suitability of the 28 prompts for inclusion in the pilot study was unanimously confirmed by the consulted clinicians.

The pilot study data collection procedure was adopted from Kocaballi et al. [9]. A team of four researchers performed the pilot study data collection. For the purposes of reliability and validity of the study, each prompt was asked to each CA or GAI model three times [9]. The responses were then recorded and transcribed. The research team also captured screen-recordings and screenshots of

Preprint JMIR Publications

visual responses.

Next, each response was assessed for suitability and prompts were excluded if they were not able to elicit a valid response from at least one CA. This resulted in three prompts being eliminated from the study, with 25 prompts deemed suitable for inclusion in the main study.

Main Study Data Collection

The main data collection process was structured into three steps, mirroring the initial three steps employed during the pilot study [9]. In other words, each of the 25 prompts was asked 3 times by the researchers. Audio responses were recorded and transcribed using Microsoft Teams, and the visual responses were screen-recorded or screen-captured.

To minimise any potential biases and confounding factors, the data collection procedures were standardised and performed under controlled conditions. Specifically, four sets of data collection (Apple Siri, Google Assistant – Pixel, ChatGPT-3.5, and BingChat) were performed on the same day, by the same researcher, within a controlled, noise-free environment. This environment was maintained consistently for all data collections. All prompts for these devices were presented by a single researcher (male, native speaker) to ensure consistency.

Two devices (Amazon Alexa and Google Assistant – Home) experienced connectivity issues and required separate sessions and therefore, adjustments in the data collection process had to be made for these specific devices. The data collection for these devices was performed elsewhere. Five devices were performed by the lead author and one device (Google Assistant – Home) was performed by another researcher on the project team.

Measurements

Assessment of Quality of Information

Quality of provided information was assessed using two widely adopted scales to measure the quality of health information collected from the internet, the JAMA benchmark criteria and the DISCERN instrument. The JAMA benchmark criteria were initially developed by Silberg et al. [35] to assess the quality of health information provided on the internet. The JAMA benchmark criteria measure information quality through the presence of four core standards: authorship, attribution, disclosure and currency, as shown in (Table 2). A score of 1 is given for the presence of each item, and 0 for the absence, with the total score ranging from 0 to 4. As it only assesses the presence of four qualities, the JAMA benchmark is the most streamlined of the widely used quality assessment tools [35, 36].

Table 2: JAMA Benchmark Criteria.

Authorship	Authors and contributors, their affiliations, and relevant credentials should be provided.
Attribution	References and sources for all content should be listed clearly, and all relevant copyright information noted.
Disclosure	Web site “ownership” should be prominently and fully disclosed, as should any sponsorship, advertising, underwriting, commercial funding arrangements or support, or potential conflicts of interest. This includes arrangements in which links to other sites are posted as a result of financial considerations. Similar standards should hold in discussion forums.

preprint JMIR Publications

Currency	Dates that content was posted and updated should be indicated.
*Each standard is scored from 0-1, overall score ranging from 0-4.	
Data source: [35]	

The DISCERN instrument is widely utilised in conjunction with the JAMA benchmark criteria to improve the rigour of the information quality assessment [37]. The DISCERN instrument was developed to enable users of health information to assess the quality of written information about treatment choices [38]. As shown in (Table 3), the instrument consists of 16 questions divided into three sections: (1) reliability, (2) quality information about treatment, and (3) an overall rating of the publication. Each question is scaled on a score of 1 to 5, with the total score ranging from 16 to 80. The scoring method of the DISCERN instrument has been slightly altered for the purposes of this study. Q2, asking about the responses' ability to achieve its aims, was skipped by many prompts due to the instructions of the DISCERN instrument. To maintain consistency across all responses, it was decided that Q2 would be omitted altogether. Furthermore, Q16 was excluded from the final score since it only consisted of the assessors' subjective rating of each response rather than the actual quality and reliability of the information provided [39]. As not all responses from this study discussed information regarding treatment of LBP, they were not assessed on section (2), thereby making those questions null for said responses. Any questions which were nulled from the instrument (Q2, 16 for all, Q9-15 for some) were excluded from the final total scores. Consistent with Weil et al., this study also evaluated the DISCERN scores into five different score categories based as shown in (Table 4) [39].

Table 3: DISCERN Instrument Questions

Q1	Are the aims clear?
Q3	Is it relevant?
Q4	Is it clear what sources of information were used to compile the publication (other than the author or producer)?
Q5	Is it clear when the information used or reported in the publication was produced?
Q6	Is it balanced and unbiased?
Q7	Does it provide details of additional sources of support and information?
Q8	Does it refer to areas of uncertainty?
Q9	Does it describe how each treatment works?
Q10	Does it describe the benefits of each treatment?
Q11	Does it describe the risks of each treatment?
Q12	Does it describe what would happen if no treatment was used?
Q13	Does it describe how treatment choices would affect overall quality of life?
Q14	Is it clear that there may be more one than possible treatment choice?
Q15	Does it provide support for shared decision making?
* (1) Q2 and Q16 have been omitted from this study as explained above. Q9-15 were omitted for responses which did not discuss treatment. (2) Each question is scored from 1-5 according to the instrument guidelines.	
Data Source: [38]	

Table 4: DISCERN Score Evaluation Scale

Score	Rating
84-100%	Excellent
68-83%	Good
52-67%	Fair
36-51%	Poor

Preprint JMIR Publications

0-35%	Very Poor
Adapted from [39]	

Assessment of Accuracy of Information

The accuracy of information was assessed by comparing the content of the responses to recommendations from international evidence-based guidelines/guidance on diagnosis and management of LBP. The selected sources included the Australian Commission on Safety and Quality in Health Care's Low Back Pain Clinical Care Standard [40], and the United Kingdom's National Institute for Health and Care Excellence (NICE) guideline on Low Back Pain and Sciatica in Over 16s: Assessment and Management [41]. These sources were selected given the quality and relevance of their information to the Australian healthcare context [42, 43]. Responses that aligned with recommendations from at least one guideline were considered accurate and given a score of 1.

Data Analysis

The data collection was primarily performed on Microsoft Teams to facilitate the recording and transcription of the responses. Each transcription was then individually checked to correct any potential errors made by the Microsoft transcription software. Transcribed data from each device was then compiled and imported into an excel document for coding. Following the process set by Weber [44], we initially defined the unit of analysis for coding as sentences within the response. This essentially refers to the idea that the researcher would scan each sentence of a response to identify common themes which fulfilled the coding scheme criteria. To ensure the reliability of coding process, a brief set of coding instructions were prepared to include both JAMA and DISCERN coding instructions for next step coding. Coding instruction for Accuracy measure was also included in this instruction.

In order to enhance the analytic rigour, two researchers used the initial coding instructions prepared earlier to independently code the responses of Alexa against the outcomes of the JAMA benchmark criteria, the DISCERN instrument and accuracy. The results from the independent coding were then compared to assess for coding process reliability. The main discrepancies between the two researchers were resolved through discussion. For instance, the researchers came to the consensus that in responses when there was one or multiple links to websites provided, only the top link was analysed as literature states that people are more inclined to click links that are ordered higher up [45]. When analysing the web pages, the researcher only analysed information provided on the page that was linked, with the exception of assessing for website disclosure as those would often be provided in an "About Us" section at the bottom of the page. Some audio-only responses mentioned a source where the information was from but did not provide a link. In these cases, the researcher did not accept it as valid authorship or referencing as it was impossible to determine which page on the website it came from or if the source is even legitimate. For reproducibility and replicability purposes, the initial coding instruction was updated with detailed consensus agreed by the researchers for the rest data coding. Appendix 2 provides the final coding instructions.

The main researcher then independently coded the rest of the data using the agreed upon coding instructions. Specifically, all three responses to each prompt by each device was coded individually by the three outcome measures. The quantitative value scores for each prompt and device were then averaged out to determine a mean score for each prompt on each device across all three measures. After which, a table was constructed for each device to break down their average scores across all three outcome measures. Tables 12-17, as seen in Appendix 2 due to space limitations, presents the

preprint JMIR Publications

average results of each of the prompts across the measures for each device.

Then, a descriptive analysis was performed on the aggregated data derived from the content analysis, followed by statistical analysis in order to identify key results. One-way ANOVA was used in conjunction with post hoc testing to assess and quantify the differences among the different devices. This was done to determine whether there were statistically significant differences between the results of the JAMA benchmark criteria and the accuracy of the CAs and GAI models. Statistical analysis was performed using SPSS software (version 28).

Results

Summary of Content Analysis

The CAs and GAI models produced 150 total responses to 25 prompts (Appendix 2). One prompt was deemed to be invalid as it did not elicit a valid response on all three times tested (P21; Google Assistant – Home). Therefore, the total number of valid responses used in the data analysis was 149.

Information Quality

Based on the information presented in Appendix 2, Tables 5, 6 and 7 summarise the results of the information quality analysis by category of prompt and device.

JAMA Benchmark Criteria

(Table 5) breaks down the mean scores of each device according to the category of prompts, as well as the overall mean score. First, three devices (Amazon Alexa, Google Assistant – Home and ChatGPT-3.5) had a mean score of 0 out of 4 in the JAMA benchmark criteria. Google Assistant – Pixel and BingChat had a mean score of 0.84 (SD 1.28) and 0.63 (SD 0.44) respectively. Siri had the highest mean score of 2.85 (SD 0.62). Across each of the categories of prompts, Siri presented the highest mean scores (range 2.43 – 3.33) compared to the other devices.

The one-way ANOVA analysis demonstrated a significant overall difference in the effects of the six devices on the JAMA score (F = 81.67, $p < 0.001$). Post hoc testing showed that significant differences were found between Siri and each of the other devices ($p < 0.001$), and between Google Assistant-Pixel and each of the other devices ($p < 0.001$), except BingChat. BingChat was found to be significantly different from Alexa, Google Assistant-home and ChatGPT ($p < 0.01$).

The breakdown of results by category of prompts is almost uniform across all four categories. Siri had the highest mean scores in each category (≥2.43), while both voice-only devices and ChatGPT-3.5 had mean scores of 0 throughout each category. Although Google Assistant – Pixel had the second highest mean scores in “Causes of LBP”, “Management and Treatment of LBP” and “Working and Exercising with LBP”, and a mean score of 0 in “Imaging of LBP”, none of the scores was higher than 1. BingChat had similar performance as Google Assistant -Pixel, but the highest score of 1 was for “Imaging of LBP” category.

Preprint JMIR Publications

Table 5: JAMA Benchmark Criteria results by category of prompt and device.

	Alexa	Google Assistant - Home*	Google Assistant - Pixel	Siri	BingChat	ChatGPT-3.5
	JAMA (0-4)	JAMA (0-4)	JAMA (0-4)	JAMA (0-4)	JAMA (0-4)	JAMA (0-4)
	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)
Causes of LBP (7)	0.00 (0)	0.00 (0)	0.86 (1.46)	2.71 (0.40)	0.48 (0.50)	0.00 (0)
Management and Treatment of LBP (9)	0.00 (0)	0.00 (0)	0.89 (1.36)	3.33 (0.60)	0.85 (0.29)	0.00 (0)
Working and Exercising with LBP (7)	0.00 (0)	0.00 (0)	1.00 (1.29)	2.43 (0.53)	0.38 (0.45)	0.00 (0)
Imaging of LBP (2)	0.00 (0)	0.00 (0)	0.00 (0)	2.67 (0.47)	1.00 (0)	0.00 (0)
Overall Score:	0.00 (0)	0.00 (0)	0.84 (1.28)	2.85 (0.62)	0.63 (0.44)	0.00 (0)

*Google Home, Working and Exercising with LBP has 6 responses as one data set is invalid, therefore the total number of responses is 149

Legend:	Voice-only	Multimodal	GAI Models

Preprint JMIR Publications

DISCERN

(Table 6) shows the mean DISCERN scores by category of prompt and device. The mean score on the DISCERN instrument across all devices was 49.33% (SD 0.14), indicating a "poor” quality evaluation assessment result. (Table 7) illustrates the breakdown of DISCERN scores by the score evaluation method adopted from Weil et al. [39]. Out of 149 total responses, only one response was classified as “excellent”, 14% (n=21) were “good” responses, 22% (n=33) were “fair” responses, 48% (n=72) were “poor” responses and 15% (n=22) were “very poor” responses.

However, each device’s performance varied. Specifically, for the prompts related to the “Causes of LBP”, the majority of responses derived from Siri were classified as “good” responses (71%). The responses of BingChat were mostly of a “fair” (43%) or “poor” quality (43%). Although Google Assistant – Pixel had some responses that were deemed as “good” (29%), most of their responses still fell into the “poor” category (71%). ChatGPT-3.5 and the two voice-only CA devices generated responses that were all either poor or very poor.

All responses from Siri to prompts regarding “Management and Treatment of LBP” fell into the “good” classification (100%). BingChat’s responses were all classified as “fair” (100%) with the majority (78%) of ChatGPT-3.5’s also having a fair evaluation. Google Assistant – Pixel was mostly poor or very poor, but some of their responses (22%) were classified as “good”. Similar to the prompts related to the “Causes of LBP”, all responses from Alexa and Google Assistant – Home fell into either the poor or very poor categories.

Regarding prompts about the ability to “Work and Exercise with LBP”, Siri had the only “excellent” response (P23), whilst most (71%) of its other responses were of a “fair” quality. Surprisingly, Google Assistant – Pixel had mostly (57%) “fair” responses. Comparatively, BingChat (71%) and ChatGPT-3.5 (100%) were mostly poor-quality responses. As consistent with other categories of prompt, Amazon Alexa and Google Assistant – Home had all responses that were either poor or very poor quality.

BingChat had all responses to prompts regarding “Imaging of LBP” that were classified as “good” quality, whilst all responses from Siri were only “fair”. The rest of the devices had all responses with “poor” information quality.

Table 6: Mean DISCERN scores by category of prompt and device

	Alexa	Google Assistant - Home	Google Assistant - Pixel	Siri	ChatGPT-3.5	BingChat	All Devices
	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)
Causes of LBP	40.14% (0.06)	36.60% (0.05)	53.06% (0.19)	69.52% (0.14)	41.09% (0.03)	48.30% (0.12)	48.12% (0.15)
Management and Treatment of LBP	33.33% (0.06)	42.54% (0.05)	49.21% (0.15)	75.56% (0.04)	54.44% (0.03)	58.78% (0.03)	52.31% (0.15)
Working and Exercising with LBP	32.86% (0.06)	40.71% (0.04)	49.18% (0.09)	62.45% (0.12)	42.04% (0.03)	52.04% (0.09)	46.69% (0.12)
Imaging of LBP	38.10% (0.01)	42.86% (0.04)	40.00% (0.04)	57.62% (0.07)	47.14% (0.02)	69.52% (0.00)	49.21% (0.12)
All Categories	35.49% (0.06)	40.38% (0.05)	49.54% (0.14)	68.76% (0.11)	46.65% (0.07)	54.82% (0.10)	49.33% (0.14)

*Google Home, Working and Exercising with LBP has 6 responses as one data set is invalid

Legend:	Very Poor	Poor	Fair	Good	Excellent

Preprint JMIR Publications

	Alexa		Google Assistant - Home*		Google Assistant - Pixel		Siri		BingChat		ChatGPT-3.5	
	Information Quality		Information Quality		Information Quality		Information Quality		Information Quality		Information Quality	
	DISCERN		DISCERN		DISCERN		DISCERN		DISCERN		DISCERN	
	Evaluation Scale	No. of Responses	Evaluation Scale	No. of Responses	Evaluation Scale	No. of Responses	Evaluation Scale	No. of Responses	Evaluation Scale	No. of Responses	Evaluation Scale	No. of Responses
Causes of LBP (7 prompts)	Very Poor	2	Very Poor	2	Very Poor	0	Very Poor	0	Very Poor	1	Very Poor	0
	Poor	5	Poor	5	Poor	5	Poor	1	Poor	3	Poor	7
	Fair	0	Fair	0	Fair	0	Fair	1	Fair	3	Fair	0
	Good	0	Good	0	Good	2	Good	5	Good	0	Good	0
	Excellent	0	Excellent	0	Excellent	0	Excellent	0	Excellent	0	Excellent	0
Management and Treatment of LBP (9 prompts)	Very Poor	7	Very Poor	1	Very Poor	2	Very Poor	0	Very Poor	0	Very Poor	0
	Poor	2	Poor	8	Poor	5	Poor	0	Poor	0	Poor	2
	Fair	0	Fair	0	Fair	0	Fair	0	Fair	9	Fair	7
	Good	0	Good	0	Good	2	Good	9	Good	0	Good	0
	Excellent	0	Excellent	0	Excellent	0	Excellent	0	Excellent	0	Excellent	0
Working and Exercising with LBP (7 prompts)	Very Poor	4	Very Poor	1	Very Poor	1	Very Poor	0	Very Poor	0	Very Poor	0
	Poor	3	Poor	5	Poor	2	Poor	0	Poor	5	Poor	7
	Fair	0	Fair	0	Fair	4	Fair	5	Fair	2	Fair	0
	Good	0	Good	0	Good	0	Good	1	Good	0	Good	0
	Excellent	0	Excellent	0	Excellent	0	Excellent	1	Excellent	0	Excellent	0
Imaging of LBP (2 prompts)	Very Poor	0	Very Poor	0	Very Poor	0	Very Poor	0	Very Poor	0	Very Poor	0
	Poor	2	Poor	2	Poor	2	Poor	0	Poor	0	Poor	2
	Fair	0	Fair	0	Fair	0	Fair	2	Fair	0	Fair	0
	Good	0	Good	0	Good	0	Good	0	Good	2	Good	0
	Excellent	0	Excellent	0	Excellent	0	Excellent	0	Excellent	0	Excellent	0

*Google Home, Working and Exercising with LBP has 6 responses as one data set is invalid, therefore the total number of responses is 149

Legend: Voice-only Multimodal GAI Models

Table 7: DISCERN score evaluation by category of prompt and device.

Preprint JMIR Publications

Preprint
JMIR Publications

Accuracy

In our investigation of the accuracy of CAs and GAI models in response to prompts regarding LBP, we observed significant variations in their performance. (Table 8) presents the accuracy of the devices based on the category of each prompt. The mean accuracy across all devices was 0.75 (SD 0.42) on a scale from 0.45 to 0.92. The voice-only based CAs scored the lowest accuracy across the board, with Alexa having a mean accuracy of 0.45 (SD 0.46) and Google Assistant – Home having a mean accuracy of 0.63 (SD 0.47). In terms of the multimodal CAs, Google Assistant – Pixel had a mean accuracy of 0.72 (SD 0.46) while Siri had a mean accuracy of 0.84 (SD 0.37). The GAI models demonstrated the highest average accuracy scores, with both BingChat and ChatGPT-3.5 having a mean accuracy of 0.92 (SD 0.28).

The one-way ANOVA analysis established significant difference in the accuracy scores across all the devices ($F = 5.59$, $p < 0.001$). Post hoc tests demonstrated significant variations between Alexa and BingChat ($p < 0.001$), Alexa and ChatGPT-3.5 ($p < 0.001$) as well as Alexa and Siri ($p < 0.05$).

In relation to the differences among all devices against each prompt category, BingChat showed the highest accuracy in response to prompts about the “Causes of LBP”, with all responses aligning with either clinical guideline. ChatGPT-3.5 and the two multimodal CAs all had a mean accuracy of 0.86 (SD 0.38). Whereas the two voice-only CAs had the lowest accuracy towards the causes of LBP, with Google Assistant – Home (0.62, SD 0.49) being slightly more accurate than Amazon Alexa (0.43, SD 0.53). GAI models are the most accurate for prompts about management and treatment of LBP, with both BingChat and ChatGPT-3.5 having a mean accuracy score of 0.89 (SD 0.33). Amazon Alexa again had the lowest accuracy with a mean accuracy of 0.44 (SD 0.53). Similar trends were seen in prompts about working and exercising with LBP, as voice-only based CAs of Amazon Alexa (0.43, SD 0.42) and Google Assistant – Home (0.52, SD 0.49) demonstrated the lowest accuracy. Siri and ChatGPT-3.5 had full accuracy scores for this category. In terms of imaging for LBP, only Amazon Alexa did not score full accuracy as it had a mean score of 0.67 (SD 0).

Table 8: Accuracy Results by Category of Prompt and Device

	Alexa	Google Assistant - Home*	Google Assistant - Pixel	Siri	BingChat	ChatGPT-3.5	All Devices
	Accuracy (0-1)	Accuracy (0-1)	Accuracy (0-1)	Accuracy (0-1)	Accuracy (0-1)	Accuracy (0-1)	Accuracy (0-1)
	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)
Causes of LBP (7)	0.43 (0.53)	0.62 (0.49)	0.86 (0.38)	0.86 (0.38)	1.00 (0)	0.86 (0.38)	0.77 (0.42)
Management and Treatment of LBP (9)	0.44 (0.53)	0.56 (0.53)	0.56 (0.53)	0.67 (0.50)	0.89 (0.33)	0.89 (0.33)	0.67 (0.48)
Working and Exercising with LBP (7)	0.43 (0.42)	0.61 (0.49)	0.71 (0.49)	1.00 (0)	0.86 (0.38)	1.00 (0)	0.77 (0.40)
Imaging of LBP (2)	0.67 (0)	1.00 (0)	1.00 (0)	1.00 (0)	1.00 (0)	1.00 (0)	0.94 (0.12)
Overall Score:	0.45 (0.46)	0.63 (0.47)	0.72 (0.46)	0.84 (0.37)	0.92 (0.28)	0.92 (0.28)	0.75 (0.42)

*Google Home, Working and Exercising with LBP has 6 responses as one data set is invalid

Legend: Voice-only Multimodal GAI Models

Preprint JMIR Publications

Discussion

In recent years, CAs and GAI models have gained significant attention and utility throughout the healthcare domain. However, our research findings showed that CAs and GAI models produced responses that were of a poor quality but moderate accuracy in response to prompts regarding LBP. Upon closer examination, this study yielded two important findings. The first finding was that Siri had the best overall performance based on a combination of information quality and accuracy score whereas voice-only CAs, Amazon Alexa and Google Assistant – Home, performed the worst in both information quality and accuracy. Secondly, GAI models had the best information accuracy but low information quality, compared to multimodal CAs.

As a multimodal device, Siri outperformed almost all other devices examined in this study, with a significantly higher mean score of 2.85 (out of 4) in the JAMA benchmark criteria compared to the voice-only CAs mean score of 0. Results from the DISCERN instrument demonstrated similar trends to that of the JAMA benchmark criteria. Our findings in relation to information quality are consistent with some of the recent studies in healthcare [9, 12]. For instance, Alagha and Helbing aimed to evaluate the quality of responses from Siri, Google Assistant and Alexa to consumer health questions about vaccines [12]. They found that Siri also scored the highest in terms of information quality measured through an adapted JAMA rubric [12]. Furthermore, our findings are also in line with those discussed by Kocabelli et al., whereby Siri had the highest number of appropriate responses and the voice-only CAs had the lowest in terms of prompts related to health and lifestyle [9]. Although Kocabelli et al. assessed the “appropriateness” of responses rather than information quality, their measure for appropriateness still evaluated the relevance of information to address the problem prompted, thus, correlated with information quality defined according to the DISCERN instrument.

The varying performance between the devices in information quality may be explained by the response style and source quality [12]. First, as Siri is a multimodal device, it supports voice and text responses thus allowing Siri to provide a brief verbal response alongside additional supporting links from webpages which contain more accurate answers. The JAMA benchmark criteria and DISCERN instrument place value upon details such as information source, and currency of data, etc. This favours the response style of Siri, as webpages often provide the necessary information required by the JAMA benchmark criteria. Although voice-only CAs often provide the cited source in their audio only response (e.g. “From livestrong.com…”), the coding scheme used in this study does not consider this as sufficient to fulfil any of the core standards of the JAMA benchmark criteria. The overall poor quality of information on treatment provided by voice-only CAs could also be explained by the limitations of providing audio only responses. When answering prompts about treatment modalities for LBP, voice-only CAs usually explicitly answer the question and do not delve further into the details of how the treatment works and how it would benefit the individual. This can be seen in Alexa’s response to P11 “From familydoctor.org, you may need one to two days of rest for a hurt back”. While primarily functioning as a voice-initiated web search, Siri can provide additional more detailed and reliable responses related to the treatments being asked, thus scoring higher [12]. A second possible explanation may be due to the different search algorithms and

Preprint JMIR Publications

prioritisation of the search results due to the different capabilities of the devices [6]. In DISCERN, most responses from Siri tended to have higher scores in regards to response relevancy, providing references and currency of information, similar to the findings of other studies [6].

Siri also performed equally as well in terms of the accuracy of its responses, with the third highest mean accuracy score and no significant difference between itself and ChatGPT-3.5 or BingChat. The voice-only CA Alexa was significantly less accurate than Siri and the GAI models, with the lowest mean accuracy score across the board. Our findings were similar to another study evaluating the COVID-19 information provided by digital voice assistants, where the voice-only CAs were identified to be less accurate than the multimodal CAs of Google Assistant – Pixel and Siri [6]. The variability in the accuracy levels of these devices may be explained by the differences in the search engines utilised by each device. Alexa uses Microsoft Bing search engine, whereas Siri and Google Assistant use Google search [46, 47]. Bing and Google have different search engine optimisation factors that influences search results [48]. This was noticeably seen in the quality of the sources of each device. Siri would usually present information from reputable sites or references backed by recognised authorities and governments or from sites that provided information largely based on expert opinion such as “Mayo Clinic” or “healthdirect.gov.au”. Whereas Alexa would provide information from much less reputable sources that were not primarily known for providing factual health information. Examples of these websites in this study included “everythingzoomer.com” or “teamsportsman.com”. Although Alexa did provide some information from reputable sources such as the ones mentioned above, most of its responses were sourced from less reputable sites. Furthermore, Alexa’s low accuracy may be due to the inherent characteristics of the device. As mentioned earlier, voice-only CAs are only able to provide audio-only responses in reply to queries. Thus, unlike the multimodal devices which are able to provide supplementary information on screen including additional webpages, the voice-only CAs were only able to vocalise their responses.

The second key finding of the study is that, compared with CAs, GAI models had high information accuracy but low information quality. ChatGPT-3.5 and BingChat had the equal highest mean accuracy score, although not significantly different from that of Siri, Google Assistant – Pixel or Google Assistant – Home. Previous studies assessing the accuracy of ChatGPT-3.5 in response to medical questions had similar findings regarding the high accuracy levels [21, 49]. For instance, Johnson et al. investigated the accuracy of ChatGPT to basic questions from almost every major medical speciality [21]. The study evaluated the responses on a 6-point Likert scale, with a score of 6 being completely accurate. The mean accuracy score of ChatGPT across 284 questions was 4.8 (between mostly and almost completely correct) [21]. Similarly, in a study investigating ChatGPT’s ability to identify the latest evidence-based practice to provide accurate responses for questions on urology, ChatGPT-3.5 had an accuracy level of 80% (24/30) [49]. The higher accuracy scores of GAI models may be attributed to the different purposes, searching sources, and technologies behind them. GAI models such as ChatGPT-3.5 are designed for general purpose, generate human-like text, trained on large collections of text data, such as books, articles, and web pages, and leveraged by the technology such as

Preprint JMIR Publications

machine learning, natural language processing, neural networks, etc. [23]. In contrast, CAs are specialised for interactive dialogues, generate human-like and contextually relevant responses, trained on finite source of large data sets of human conversation, user queries and responses.

Although performing well in accuracy, none of ChatGPT-3.5's responses fulfilled any of the core standards of JAMA, whereas the only core standard that BingChat would occasionally display was "attribution". This is similar to the information quality assessment in other literature [50]. These results are an indictment of the response style of GAI models. Due to the nature of the models generating the response rather than presenting data taken from other sources, answers generated through ChatGPT-3.5 do not provide references. Thus, it is unsurprising that it scored zero on every response regarding each of the core standards of the JAMA benchmark criteria. BingChat, on the other hand, differs from ChatGPT-3.5 in this sense as it would sometimes supply readers with the relevant sources of information, as well as supplementary information that could be useful. In terms of the DISCERN score, although GAI models overall had lower information quality compared to some CAs, they were comparatively even when only assessing section two questions (Quality Information about Treatments). Furthermore, when responding to prompts about "Management and Treatment of LBP", both GAI models had a mostly "fair" information quality. This may partially be due to the GAI models providing more accurate and details answers. The length of the responses of GAI devices was also far greater than those of CAs, thus allowing GAI models to go into far greater detail in response to prompts about treatment. In addition, ChatGPT-3.5 repeatedly emphasised the importance of patients discussing their concerns about LBP with a healthcare professional and encouraged shared decision making. This is significant as it allowed ChatGPT-3.5's responses to score 5 (out of 5) on Q15 of the DISCERN instrument every time that quality information about treatments was applicable. As a result, the main downfall of information quality for GAI models is rather in the reliability of information.

Preprint JMIR Publications

Theoretical and Practical Implications

This study highlights a concerning picture of the quality and accuracy of information provided by CAs regarding LBP. Consistent with prior CAs studies on other healthcare topics, we found a major variation in information quality and accuracy for LBP patients. Thus, this study contributes to existing LBP literature and advances our understanding of CAs' role in healthcare information retrieval. In addition, this study is amongst the first to explore the effectiveness of GAI model's in assisting LBP patients to find high quality and accurate information. Finally, the 25 prompts developed in this study can also be used by other researchers in the future as they were derived through a thorough literature and validated through the pilot study and 149 responses in main study.

Considering these findings, this study has three practical implications. Firstly, clinicians should be cautious when recommending the use of CAs and GAI for LBP patients. Information from CAs, especially from voice-only devices, should always be supplemented with more in-depth information from up-to-date, professional, and evidence-based resources, especially for treatment and exercise related recommendations. Secondly, given the better performance of multimodal than voice-only interface, developers of future health care CAs may consider using multimodal device for quality assurance purposes. Thirdly, as GAI models can provide high accurate information even just with single prompt, we may capture the full potential of the GAI models if the correct follow up prompts can be provided to the GAI models, which is often the case in real-life with patients and clinicians having follow-up questions to clarify. Thus, patients may be recommended to use suitable follow-up prompts with GAI models in order to have a better understanding of the LBP related questions.

Limitations and Future Directions

There are several limitations of the study which need to be acknowledged. Firstly, the data was collected on devices that were not in their factory settings. Considering that the researchers either work or study in health-related fields, it is possible that their previous search history and activities with their devices may have affected the responses they received for health-related questions compared to the lay individual, therefore affecting the generalisability of the findings. Future research in this area should aim to use devices that are in their factory settings in order to remove any external variables that may affect any results. Secondly, the selection of the outcome measures, namely the JAMA benchmark criteria and DISCERN instrument, may introduce bias into the assessment of the devices. These outcome measures were primarily developed for evaluating health information on websites, thus making them more suitable for devices that provide website links in their responses. This may disproportionately favour devices that rely on external web sources to answer questions, potentially skewing the assessment results towards multimodal CAs and GAI models. This could be addressed in the future by developing a valid and reliable measurement tool to assess CAs and GAI models in response to questions about health [12]. However, one potential issue of developing a scoring rubric for a study would be the loss of standardisation of the measurement tools, in which case, inter-rater reliability would be required to ensure that the measurements are reliable and valid. Thirdly, this research study utilised a binary coding scheme to measure accuracy, where the responses were either coded as accurate (scored with 1) or inaccurate (scored with 0). This may potentially oversimplify the assessment of accuracy of the responses, and potentially disregard nuances in the correctness of information provided. It did not account for answers that were partially accurate, therefore potentially leading to an incomplete and simplified assessment of the device’s performances. A more detailed scale, similar to the one used in Johnson et al., may be required for a better understanding of the accuracy of the CAs and GAI models [21]. Lastly, the rapidly evolving nature of AI models may render the findings of this study to be outdated over time. The companies that produce the commonly used CAs and GAI models which have been examined in this study are constantly updating not only their hardware but also the software, potentially causing the AI models assessed in this research study to become obsolete as newer models with improved functionality emerge after the conclusion of the study. Just recently, OpenAI announced that ChatGPT will be able to search real time data to answer questions with the latest available information on any topic [51]. Therefore, future research is needed to assess the information quality and accuracy of the updates devices

Preprint JMIR Publications

and software's to ensure that new recommendations can be made.

Conclusion

This study found that there was a large variation in the information quality and accuracy of different CAs responding to prompts about LBP. We also found that Apple's Siri on the iPhone was the best CA in providing patients with both high quality and accurate responses to prompts regarding causes, management and treatment, the ability to work or exercise and imaging of LBP. Furthermore, GAI models were found to be highly accurate but of low information quality when responding to prompts about LBP. However, due to the highly variable nature of LBP cases, CAs and GAI models should not become a replacement for seeing a health professional. Rather these findings will be able to guide clinicians in recommending patients devices where they can most easily find reliable and accurate information to help them self-manage their LBP.

Acknowledgements

MLF is supported by the National Health and Medical Research Council of Australia Leadership Fellowship

Conflicts of Interest

None declared.

Abbreviations

CA: conversational agent
GAI: generative artificial intelligence
GPT: generative pretrained transformer
JMIR: Journal of Medical Internet Research
LBP: low back pain

Preprint JMIR Publications

References

1. Hartvigsen, J., et al., *What low back pain is and why we need to pay attention.* Lancet, 2018. **391**(10137): p. 2356-2367.
2. Australian Institute of Health and Welfare, *Australian burden of disease study 2022*. 2022, AIHW: Canberra.
3. Ferreira, M.L., et al., *Global, regional, and national burden of low back pain, 1990-2020, its attributable risk factors, and projections to 2050: a systematic analysis of the Global Burden of Disease Study 2021.* Lancet Rheumatol, 2023. **5**(6): p. e316-e329.
4. Australian Institute of Health and Welfare, *Back problems*. 2020, AIHW: Canberra.
5. Foster, N.E., et al., *Prevention and treatment of low back pain: evidence, challenges, and promising directions.* Lancet, 2018. **391**(10137): p. 2368-2383.
6. Goh, A., L. Wong, and K. Yap, *Evaluation of COVID-19 information provided by digital voice assistants.* International Journal of Digital Health, 2021. **1**: p. 3.
7. Owens, O.L., M. Leonard, and A. Singh, *Efficacy of Alexa, Google Assistant, and Siri for supporting informed prostate cancer screening decisions for african-american men.* J Cancer Educ, 2023. **38**(5): p. 1752-1759.
8. Alnefaie, A., et al. *An overview of conversational agent: applications, challenges and future directions*. in *WEBIST*. 2021.
9. Kocaballi, A.B., et al., *Responses of conversational agents to health and lifestyle prompts: Investigation of appropriateness and presentation structures.* J Med Internet Res, 2020. **22**(2): p. e15823.
10. Tudor Car, L., et al., *Conversational agents in health care: scoping review and conceptual analysis.* J Med Internet Res, 2020. **22**(8): p. e17158.
11. Allouch, M., A. Azaria, and R. Azoulay, *Conversational agents: goals, technologies, vision and challenges.* Sensors (Basel), 2021. **21**(24).
12. Alagha, E.C. and R.R. Helbing, *Evaluating the quality of voice assistants' responses to consumer health questions about vaccines: an exploratory comparison of Alexa, Google Assistant and Siri.* BMJ Health Care Inform, 2019. **26**(1).
13. Laranjo, L., et al., *Conversational agents in healthcare: A systematic review.* J Am Med Inform Assoc, 2018. **25**(9): p. 1248-1258.
14. Merino, J., et al., *A data quality in use model for big data.* Future Gener Comput Syst, 2016. **63**: p. 123-130.
15. Bailey, J.E. and S.W. Pearson, *Development of a tool for measuring and analyzing computer user satisfaction.* Management Science, 1983. **29**(5): p. 530-545.
16. Miner, A.S., et al., *Smartphone-based conversational agents and responses to questions about mental health, interpersonal violence, and physical health.* JAMA Intern Med, 2016. **176**(5): p. 619-625.
17. Fabio Duarte. *Number of ChatGPT users*. 2023; Available from: https://explodingtopics.com/blog/chatgpt-users.

18. Elton Grivith Dsouza. *How ChatGPT works: Training model of ChatGPT,;*. 2023; Available from: https://www.edureka.co/blog/how-chatgpt-works-training-model-of-chatgpt/.
19. OpenAI. *Introducing ChatGPT*. OpenAI 2022; Available from: https://openai.com/blog/chatgpt.
20. Baidoo-Anu, D. and L.O. Ansah, *Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning.* Journal of AI, 2023. **7**(1): p. 52-62.
21. Johnson, D., et al., *Assessing the accuracy and reliability of AI-generated medical responses: an evaluation of the Chat-GPT model.* Res Sq, 2023.
22. Lombardo, R. and C. De Nunzio, *Nomograms in PCa: where do we stand.* Prostate Cancer Prostatic Dis, 2023. **26**(3): p. 447-448.
23. Cocci, A., et al., *Quality of information and appropriateness of ChatGPT outputs for urology patients.* Prostate Cancer Prostatic Dis, 2023.
24. Dingler, T., et al., *The use and promise of conversational agents in digital health.* Yearb Med Inform, 2021. **30**(01): p. 191-199.
25. Krippendorff, K., *Content analysis: An introduction to its methodology*. 2018: Sage publications.
26. Hoy, M.B., *Alexa, Siri, Cortana, and more: an introduction to voice assistants.* Med Ref Serv Q, 2018. **37**(1): p. 81-88.
27. Snead, A. *What search engine does Alexa use? And can I use Google to…*. Smarter Home Guides 2020 05/10/2023; Available from: https://smarterhomeguide.com/alexa-search-engine/.
28. Google. *How Google Assistant helps you get things done*. 2023; Available from: https://developers.google.com/assistant/how-assistant-works.
29. Balakrishnan, A. *Google is becoming the default search engine for more Apple products, striking a deal that could be worth billions*. CNBC 2017; Available from: https://www.cnbc.com/2017/09/25/siri-search-results-switch-to-google-from-bing.html.
30. Mehdi, Y. *Reinventing search with a new AI-Powered Microsoft Bing and Edge, Your copilot for the web*. Official Microsoft Blog 2023; Available from: https://blogs.microsoft.com/blog/2023/02/07/reinventing-search-with-a-new-ai-powered-microsoft-bing-and-edge-your-copilot-for-the-web/.
31. Mearian, L. *What are LLMs, and how are they used in Generative AI?* Computerworld 2023; Available from: https://www.computerworld.com/article/3697649/what-are-large-language-models-and-how-are-they-used-in-generative-ai.html.
32. Microsoft. *What is BingChat, and how can you use it?* 2023; Available from: https://www.microsoft.com/en-us/bing/do-more-with-ai/what-is-bing-chat-and-how-can-you-use-it.
33. Coole, C., et al., *What concerns workers with low back pain? Findings of a qualitative study of patients referred for rehabilitation.* J Occup Rehabil, 2010. **20**(4): p. 472-480.
34. Australian Commission on Safety and Quality in Health Care, *Common questions about low back pain - information for patients*, ACSQH, Editor. 2022.
35. Silberg, W.M., G.D. Lundberg, and R.A. Musacchio, *Assessing, controlling, and*

assuring the quality of medical information on the internet: Caveant lector et viewor—Let the reader and viewer beware. JAMA, 1997. **277**(15): p. 1244-1245.

36. Cassidy, J.T. and J.F. Baker, *Orthopaedic patient information on the World Wide Web: An essential review.* J Bone Joint Surg Am, 2016. **98**(4): p. 325-38.
37. Eksi Ozsoy, H., *Evaluation of YouTube videos about smile design using the DISCERN tool and Journal of the American Medical Association benchmarks.* J Prosthet Dent, 2021. **125**(1): p. 151-154.
38. Charnock, D., et al., *DISCERN: an instrument for judging the quality of written consumer health information on treatment choices.* J Epidemiol Community Health, 1999. **53**(2): p. 105-111.
39. Weil, A.G., et al., *Evaluation of the quality of information on the Internet available to patients undergoing cervical spine surgery.* World Neurosurg, 2014. **82**(1-2): p. e31-e39.
40. Australian Commission on Safety and Quality in Health Care, *Low Back Pain Clinical Care Standard*. 2022: ACSQH.
41. National Institute for Health and Care Excellence, *National Institute for Health and Care Excellence: Guidelines*, in *Low back pain and sciatica in over 16s: Assessment and management*. 2016, National Institute for Health and Care Excellence (NICE): London.
42. Castellini, G., et al., *Are clinical practice guidelines for low back pain interventions of high quality and updated? A systematic review using the AGREE II instrument.* BMC Health Serv Res, 2020. **20**(1): p. 970.
43. Doniselli, F.M., et al., *A critical appraisal of the quality of low back pain practice guidelines using the AGREE II tool and comparison with previous evaluations: a EuroAIM initiative.* Eur Spine J, 2018. **27**(11): p. 2781-2790.
44. Weber, R.P., *Basic content analysis*. Vol. 49. 1990: Sage.
45. Keane, M.T., M. O'Brien, and B. Smyth, *Are people biased in their use of search engines?* Commun ACM, 2008. **51**(2): p. 49-52.
46. Tom Warren. *Microsoft and Amazon partner to integrate Alexa and Cortana digital assistants*. The Verge 2017; Available from: https://www.theverge.com/2017/8/30/16224876/microsoft-amazon-cortana-alexa-partnership.
47. Matthew Panzarino. *Apple switches from Bing to Google for Siri web search results on iOS and Spotlight for Mac*. TechCrunch 2017; Available from: https://techcrunch.com/2017/09/25/apple-switches-from-bing-to-google-for-siri-web-search-results-on-ios-and-spotlight-on-mac/.
48. Theuring, J. and K. Barnard. *Bing vs Google: Search engine comparison 2023*. Impression Digital 2023 17/08/2023; Available from: https://www.impressiondigital.com/blog/bing-differ-google/.
49. Zhou, Z., et al., *Is ChatGPT an evidence-based doctor?* Eur Urol, 2023. **84**(3): p. 355-356.
50. Hurley, E.T., et al., *Evaluation high-quality of information from ChatGPT (Artificial Intelligence—Large Language Model) Artificial Intelligence on shoulder stabilization surgery.* Arthroscopy, 2023.
51. Gold, J. *ChatGPT can now look at the web - for real this time*. Computerworld

Preprint
JMIR Publications

2023; Available from: https://www.computerworld.com/article/3709189/chatgpt-can-now-look-at-the-web-for-real-this-time.html.

52. DISCERN Online. *DISCERN Instrument general instructions*. Available from: http://www.discern.org.uk/discern_instrument.php.

Preprint
JMIR Publications

Appendices

Appendix 1: Coding instructions.

Table 9: DISCERN rubric*

	Q	1	2	3	4	5
Q1	Are the aims clear?	No aims mentioned	Mentions what the response entails	Mentions goals of the response	Mentions goals of the response and what it entails	Mentions goals of the response, what it entails and who might find it useful
Q3	Is it relevant?	The information is not relevant to the question at all	The information is slightly relevant to the question	The information is partially relevant to the question	The information is mostly relevant to the question	The information provided is relevant to the question asked
Q4	Is it clear what sources of information were used to compile the publication?	No references and no bibliography	Some intext citations and no bibliography	In text citations for most claims but no bibliography	In text citation and bibliography for some claims	In text citations and bibliography for most claims
Q5	Is it clear when the information used or reported in the publication was produced?	No date of when the article/response was produced	Dates of some of the sources and no date for the whole article	Date of when the article was posted only	Date of the article, and some dates of sources	Date of when the article was posted, as well as any revisions or updates to the publication and dates of when sources were published

Preprint JMIR Publications

Q6	Is it balanced and unbiased?	The information is completely unbalanced and biased	The information is partially biased	The response is written from an objective perspective	The response is written from an objective perspective, but it only has either a range of supporting sources or an external assessment	The response is written from an objective perspective, a range of supporting sources have been used to write it and there has been an external assessment of the response or article
Q7	Does it provide details of additional support and information?	No additional information has been provided	One additional source provided	Two or more additional sources provided	Two or more additional sources provided, with description	The response provides multiple additional sources or pages of information
Q8	Does it refer to areas of uncertainty?	No uncertainty about treatment choices is mentioned	Some uncertainty is mentioned about treatment choices but no extra discussion is made about it		Some uncertainty is mentioned about treatment choices or other information	The response makes a clear reference to any uncertainty regarding treatment choices: may be linked to each treatment choice or be covered

Preprint JMIR Publications

						in a more general discussion or summary of the choices mentioned in the publication
Q9	Does it describe how each treatment works?	None of the descriptions about treatments explains how it works	A few of the descriptions about treatments explain how it works	Some of the descriptions about treatments explains how it works	Most of the descriptions about treatments explains how it works	A description of how each treatment includes details of how it works
Q10	Does it describe the benefits of each treatment?	No benefits are described for any treatments	A few of the descriptions about treatments explain the benefits of doing it	Some benefits are described for some of the treatments	Most treatments describe have a benefit included	A benefit is described for each treatment
Q11	Does it describe the risks of each treatment	No risks are described for any of the treatments	A risk is described for a few of the treatments	A risk is described for some of the treatments	A risk is described for most if not all treatments	A risk is described for most if not all treatments, and how it can be avoided
Q12	Does it describe what would happen if no treatment was used?	The publication does not include any reference to the risks or benefits of not undertaking a specific treatment	A description of either the risks or the benefits of not performing a specific treatment			There is a clear description of a risk or a benefit associated with any no treatment option (in the case of

Preprint JMIR Publications

						LBP, usually referring to the natural course of LBP)
Q13	Does it describe how the treatment will affect overall quality of life?	There is no reference to overall quality of life in relation to treatment choices	The response includes some references to overall quality of life in relation to treatment choices, but the information is unclear or incomplete			The response includes a clear reference to overall quality of life, especially activity limitations, participation restrictions and functional capacity, in relation to any of the treatment choices mentioned
Q14	Is it clear that there may be more than one possible treatment choice	The response does not give any indication that there may be a choice about treatment	The response indicates that there may be more than one treatment choice, but the information is unclear or incomplete			The response indicates that there may be more than one treatment choice, but the information is unclear or incomplete
Q15	Does it provide support for	The response does not provide any	The response mentions	The response mentions	The response actively	The response mentions

Preprint JMIR Publications

	shared decision making?	support or mention of shared decision making	shared decision making	the different options for shared decision making	encourages the patient to engage in shared decision making	the different options for shared decision making and actively encourages the patient to engage in it
Adapted from [52], *Omits Q2 and Q16						

Preprint JMIR Publications

Table 10: JAMA Benchmark criteria rubric

Core Standard	Criteria for presence
Authorship	Authors and contributors, their affiliations, and relevant credentials are present.
Attribution	References and sources for all content is listed clearly, and all relevant copyright information noted.
Disclosure	Website “ownership” is prominently and fully disclosed.
Currency	Dates that content as posted and updated are indicated.
Adapted from [35]	

Table 11: Key agreements for coding analysis

When a response provides two or more links, which ones should be analysed?	Only the top link should be analysed.
When a response provides links to a specific page on a website, what information should be analysed?	Only the information on the page should be analysed, with exception for the “About Us” page as it may provide key information about website disclosure.
If an audio response mentions a website that it got its information from, should that be counted as a valid source or website?	If no link was provided in the response, then it would not be accepted as a reference or source.

Preprint JMIR Publications

Appendix 2: Tables of coding result analysis from all devices.

Table 12: Alexa coding and analysis results.

Category of Prompt	Prompt No.	JAMA Average	DISCERN Average	DISCERN Total	Percentage	Accuracy Average
Causes of LBP	P1	0	16.00	35	45.71%	1.00
	P2	0	13.67	35	39.05%	1.00
	P3	0	13.00	35	37.14%	0.00
	P4	0	11.67	35	33.33%	0.00
	P5	0	17.00	35	48.57%	1.00
	P6	0	15.00	35	42.86%	0.00
	P7	0	12.00	35	34.29%	0.00
Management and Treatment of LBP	P8	0	30.00	70	42.86%	0.00
	P9	0	25.00	70	35.71%	0.00
	P10	0	20.00	70	28.57%	0.00
	P11	0	20.00	70	28.57%	1.00
	P12	0	22.00	70	31.43%	1.00
	P13	0	24.00	70	34.29%	1.00
	P14	0	19.00	70	27.14%	0.00
	P15	0	30.00	70	42.86%	0.00
	P16	0	20.00	70	28.57%	1.00
Working and Exercising with LBP	P17	0	23.33	70	33.33%	0.67
	P18	0	14.33	70	20.48%	0.67
	P19	0	23.00	70	32.86%	0.00
	P20	0	13.00	35	37.14%	0.00
	P21	0	27.67	70	39.52%	0.67
	P22	0	11.67	35	33.33%	0.00
	P23	0	11.67	35	33.33%	1.00
Imaging	P24	0	13.00	35	37.14%	0.67
	P25	0	13.67	35	39.05%	0.67

Table 13: BingChat coding and analysis results.

Category of Prompt	Prompt No.	JAMA Average	DISCERN Average	DISCERN Total	Percentage	Accuracy Average
Causes of LBP	P1	1	21.67	35	61.90%	1.00
	P2	0	14.67	35	41.90%	1.00
	P3	1	22.00	35	62.86%	1.00
	P4	0.33	15.33	35	43.81%	1.00
	P5	1	20.00	35	57.14%	1.00
	P6	0	12.00	35	34.29%	1.00
	P7	0	12.67	35	36.19%	1.00
Management and Treatment of LBP	P8	1	41.33	70	59.05%	1.00
	P9	1	40.00	70	57.14%	1.00
	P10	0.33	38.00	70	54.29%	1.00
	P11	1	43.67	70	62.38%	1.00
	P12	1	43.00	70	61.43%	1.00
	P13	1	41.00	70	58.57%	1.00
	P14	1	44.00	70	62.86%	1.00
	P15	0.33	37.67	70	53.81%	0.00
	P16	1	41.67	70	59.52%	1.00
Working and Exercising with LBP	P17	0.33	16.33	35	46.67%	1.00
	P18	0	35.33	70	50.48%	1.00
	P19	1	22.33	35	63.81%	1.00
	P20	1	23.00	35	65.71%	1.00
	P21	0	31.00	70	44.29%	0.00
	P22	0	15.00	35	42.86%	1.00
	P23	0.33	17.67	35	50.48%	1.00
Imaging	P24	1	24.33	35	69.52%	1.00
	P25	1	24.33	35	69.52%	1.00

Table 14: ChatGPT-3.5 coding and analysis results

Category of Prompt	Prompt No.	JAMA Average	DISCERN Average	DISCERN Total	Percentage	Accuracy Average
Causes of LBP	P1	0.00	14.00	35	40.00%	1.00
	P2	0.00	13.33	35	38.10%	1.00
	P3	0.00	14.33	35	40.95%	0.00
	P4	0.00	13.67	35	39.05%	1.00
	P5	0.00	16.33	35	46.67%	1.00
	P6	0.00	15.00	35	42.86%	1.00
	P7	0.00	14.00	35	40.00%	1.00
Management and Treatment of LBP	P8	0.00	37.00	70	52.86%	1.00
	P9	0.00	36.00	70	51.43%	1.00
	P10	0.00	39.00	70	55.71%	1.00
	P11	0.00	39.00	70	55.71%	1.00
	P12	0.00	41.00	70	58.57%	1.00
	P13	0.00	34.00	70	48.57%	1.00
	P14	0.00	40.00	70	57.14%	1.00
	P15	0.00	39.67	70	56.67%	0.00
	P16	0.00	37.33	70	53.33%	1.00
Working and Exercising with LBP	P17	0.00	14.00	35	40.00%	1.00
	P18	0.00	16.00	35	45.71%	1.00
	P19	0.00	14.00	35	40.00%	1.00
	P20	0.00	15.00	35	42.86%	1.00
	P21	0.00	13.00	35	37.14%	1.00
	P22	0.00	16.00	35	45.71%	1.00
	P23	0.00	15.00	35	42.86%	1.00
Imaging	P24	0.00	17.00	35	48.57%	1.00
	P25	0.00	16.00	35	45.71%	1.00

Preprint JMIR Publications

Table 15: Google Assistant – Home coding and analysis results.

Category of Prompt	Prompt No.	JAMA Average	DISCERN Average	DISCERN Total	Percentage	Accuracy Average
Causes of LBP	P1	0.00	13.00	35	37.14%	1.00
	P2	0.00	9.00	35	25.71%	0.00
	P3	0.00	15.00	35	42.86%	0.00
	P4	0.00	13.67	35	39.05%	1.00
	P5	0.00	14.00	35	40.00%	1.00
	P6	0.00	13.00	35	37.14%	1.00
	P7	0.00	12.00	35	34.29%	0.33
Management and Treatment of LBP	P8	0.00	30.00	70	42.86%	0.00
	P9	0.00	25.00	70	35.71%	0.00
	P10	0.00	32.00	70	45.71%	1.00
	P11	0.00	30.00	70	42.86%	1.00
	P12	0.00	34.00	70	48.57%	1.00
	P13	0.00	27.00	70	38.57%	1.00
	P14	0.00	25.67	70	36.67%	0.00
	P15	0.00	29.67	70	42.38%	0.00
	P16	0.00	34.67	70	49.52%	1.00
Working and Exercising with LBP	P17	0.00	13.33	35	38.10%	1.00
	P18	0.00	30.00	70	42.86%	1.00
	P19	0.00	28.33	70	40.48%	0.00
	P20	0.00	12.00	35	34.29%	0.00
	P21	NA	NA	NA	NA	NA
	P22	0.00	16.00	35	45.71%	1.00
	P23	0.00	15.00	35	42.86%	0.67
Imaging	P24	0.00	14.00	35	40.00%	1.00
	P25	0.00	16.00	35	45.71%	1.00

Preprint
JMIR Publications

Table 16: Google Assistant – Pixel coding and analysis results.

Category of Prompt	Prompt No.	JAMA Average	DISCERN Average	DISCERN Total	Percentage	Accuracy Average
Causes of LBP	P1	0.00	15.00	35	42.86%	1.00
	P2	3.00	57.00	70	81.43%	1.00
	P3	3.00	55.00	70	78.57%	1.00
	P4	0.00	16.00	35	45.71%	1.00
	P5	0.00	16.00	35	45.71%	1.00
	P6	0.00	14.00	35	40.00%	1.00
	P7	0.00	13.00	35	37.14%	0.00
Management and Treatment of LBP	P8	0.00	31.00	70	44.29%	0.00
	P9	3.00	51.00	70	72.86%	1.00
	P10	3.00	53.00	70	75.71%	1.00
	P11	0.00	24.00	70	34.29%	1.00
	P12	0.00	30.00	70	42.86%	1.00
	P13	2.00	34.00	70	48.57%	0.00
	P14	0.00	29.00	70	41.43%	0.00
	P15	0.00	24.00	70	34.29%	0.00
	P16	0.00	34.00	70	48.57%	1.00
Working and Exercising with LBP	P17	0.00	30.00	70	42.86%	1.00
	P18	0.00	31.00	70	44.29%	1.00
	P19	0.00	43.00	70	61.43%	0.00
	P20	3.00	37.00	70	52.86%	1.00
	P21	2.00	38.00	70	54.29%	1.00
	P22	0.00	12.00	35	34.29%	0.00
	P23	2.00	19.00	35	54.29%	1.00
Imaging	P24	0.00	13.00	35	37.14%	1.00
	P25	0.00	15.00	35	42.86%	1.00

Preprint JMIR Publications

Table 17: Siri coding and analysis results.

Category of Prompt	Prompt No.	JAMA Average	DISCERN Average	DISCERN Total	Percentage	Accuracy Average
Causes of LBP	P1	3.00	53.00	70	75.71%	1.00
	P2	3.00	57.00	70	81.43%	1.00
	P3	3.00	55.00	70	78.57%	1.00
	P4	2.00	49.00	70	70.00%	1.00
	P5	3.00	15.00	35	42.86%	0.00
	P6	2.33	21.33	35	60.95%	1.00
	P7	2.67	27.00	35	77.14%	1.00
Management and Treatment of LBP	P8	2.33	50.33	70	71.90%	1.00
	P9	3.00	48.00	70	68.57%	1.00
	P10	3.00	53.00	70	75.71%	0.00
	P11	4.00	56.00	70	80.00%	1.00
	P12	3.67	56.67	70	80.95%	1.00
	P13	4.00	57.00	70	81.43%	0.00
	P14	3.00	51.00	70	72.86%	1.00
	P15	4.00	51.00	70	72.86%	0.00
	P16	3.00	53.00	70	75.71%	1.00
Working and Exercising with LBP	P17	2.00	42.00	70	60.00%	1.00
	P18	2.00	39.00	70	55.71%	1.00
	P19	3.00	50.00	70	71.43%	1.00
	P20	3.00	37.00	70	52.86%	1.00
	P21	2.00	38.00	70	54.29%	1.00
	P22	2.00	20.00	35	57.14%	1.00
	P23	3.00	30.00	35	85.71%	1.00
Imaging	P24	3.00	22.00	35	62.86%	1.00
	P25	2.33	18.33	35	52.38%	1.00

Preprint JMIR Publications

Appendix 3: Prompts.

Table 18: Prompts in main study.

P1	What are the causes of my low back pain?
P2	Is my low back pain caused by specific injury or accident?
P3	Is my low back pain due to work?
P4	Is my low back pain due to a medical condition?
P5	How do I know if my low back pain is due to a muscle strain or something more serious?
P6	Is my low back pain due to my weight or level of fitness?
P7	How can I determine the exact cause of my low back pain?
P8	What are effective treatments of my low back pain?
P9	How can I manage my low back pain at home?
P10	What medications should I take for my low back pain?
P11	How long should I rest if I'm experiencing low back pain?
P12	Should I go to physical therapy for my low back pain?
P13	What exercises or stretches should I do for my low back pain?
P14	Should I go to a chiropractor for my low back pain?
P15	Should I use heat or ice for my low back pain?
P16	Do I need surgery for my low back pain?
P17	Can I keep working with low back pain?
P18	Can I do physical activity or exercise with my low back pain?
P19	Are there any specific exercises that will make my back worse?
P20	Are there any specific work activities that will make my back worse?

Preprint JMIR Publications

P21	How should I change my work or physical activities to avoid making my back worse?
P22	Can I keep playing sports with low back pain?
P23	Is there a risk of further injury to my back if I do physical activity with low back pain?
P24	Do I need an X-Ray or MRI for low back pain?
P25	What will an X-Ray tell me about my low back pain?

Preprint JMIR Publications

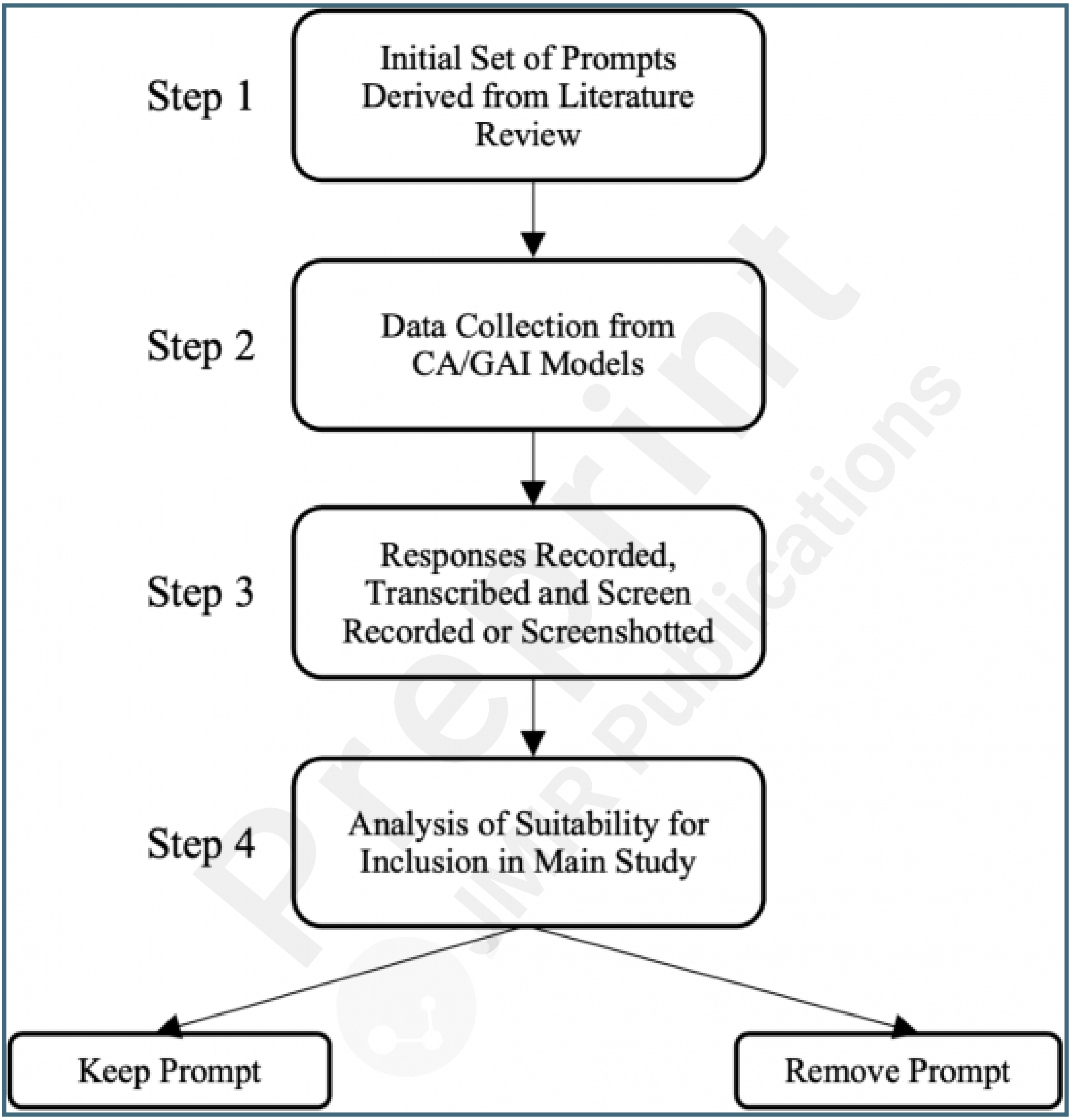
Supplementary Files

Preprint
JMIR Publications

Figures

Preprint JMIR Publications

Flow chart of pilot study process.


Step 1
Initial Set of Prompts Derived from Literature Review
Step 2
Data Collection from CA/GAI Models
Step 3
Responses Recorded, Transcribed and Screen Recorded or Screenshotted
Step 4
Analysis of Suitability for Inclusion in Main Study
Keep Prompt
Remove Prompt

Preprint JMIR Publications