Original Paper

ChatGPT vs Google for Queries Related to Dementia and Other Cognitive Decline: Comparison of Results

Vagelis Hristidis[1], PhD; Nicole Ruggiano[2], MSW, PhD; Ellen L Brown[3], RN, MS, EdD; Sai Rithesh Reddy Ganta[1], BSc; Selena Stewart[2]

[1]Department of Computer Science and Engineering, University of California, Riverside, Riverside, CA, United States

[2]School of Social Work, University of Alabama, Tuscaloosa, AL, United States

[3]Nicole Wertheim College of Nursing and Health Sciences, Florida International University, Miami, FL, United States

Corresponding Author:
Vagelis Hristidis, PhD
Department of Computer Science and Engineering
University of California, Riverside
Winston Chung Hall, Room 317
Riverside, CA, 92521
United States
Phone: 1 9518272478
Email: vagelis@cs.ucr.edu

Abstract

Background: People living with dementia or other cognitive decline and their caregivers (PLWD) increasingly rely on the web to find information about their condition and available resources and services. The recent advancements in large language models (LLMs), such as ChatGPT, provide a new alternative to the more traditional web search engines, such as Google.

Objective: This study compared the quality of the results of ChatGPT and Google for a collection of PLWD-related queries.

Methods: A set of 30 informational and 30 service delivery (transactional) PLWD-related queries were selected and submitted to both Google and ChatGPT. Three domain experts assessed the results for their currency of information, reliability of the source, objectivity, relevance to the query, and similarity of their response. The readability of the results was also analyzed. Interrater reliability coefficients were calculated for all outcomes.

Results: Google had superior currency and higher reliability. ChatGPT results were evaluated as more objective. ChatGPT had a significantly higher response relevance, while Google often drew upon sources that were referral services for dementia care or service providers themselves. The readability was low for both platforms, especially for ChatGPT (mean grade level 12.17, SD 1.94) compared to Google (mean grade level 9.86, SD 3.47). The similarity between the content of ChatGPT and Google responses was rated as high for 13 (21.7%) responses, medium for 16 (26.7%) responses, and low for 31 (51.6%) responses.

Conclusions: Both Google and ChatGPT have strengths and weaknesses. ChatGPT rarely includes the source of a result. Google more often provides a date for and a known reliable source of the response compared to ChatGPT, whereas ChatGPT supplies more relevant responses to queries. The results of ChatGPT may be out of date and often do not specify a validity time stamp. Google sometimes returns results based on commercial entities. The readability scores for both indicate that responses are often not appropriate for persons with low health literacy skills. In the future, the addition of both the source and the date of health-related information and availability in other languages may increase the value of these platforms for both nonmedical and medical professionals.

(J Med Internet Res 2023;25:e48966) doi: 10.2196/48966

KEYWORDS

chatbots; large language models; ChatGPT; web search; language model; Google; aging; cognitive; cognition; dementia; gerontology; geriatric; geriatrics; query; queries; information seeking; search

Introduction

Background

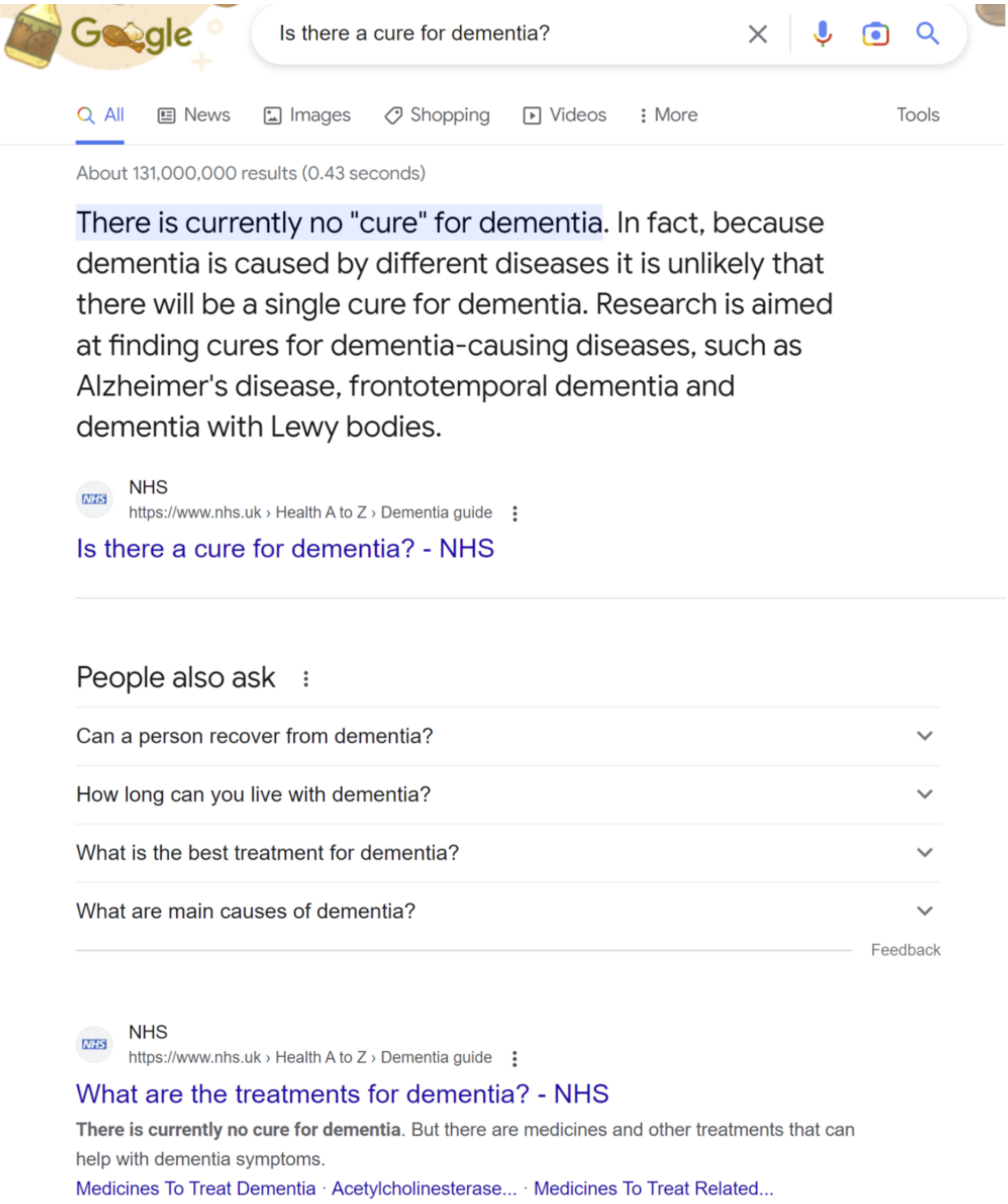
People living with dementia or other cognitive decline and their caregivers (collectively referred to as PLWD in this paper) often find it challenging to obtain the right information about their health, given the wide range of conditions, progression levels, symptoms, and side effects associated with dementia or other cognitive decline [1,2]. These problems are exacerbated due to the shortage of expert providers and the limited access to them, especially in rural or low-income settings [3,4]. PLWD have been using the web to find answers, often starting from a web search engine and then following relevant hyperlinks [3,4]. This is challenging, given the information overload on the web, the reliability of sources, the health literacy level of some content, and the skill required to locate the right answer [5]. Another factor that makes such query tools challenging is the wide range of education levels of PLWD.

Recent advances in deep learning, which is a type of machine learning based on artificial neural networks, have given rise to several conversational artificial intelligence (AI) platforms (chatbots). In our recent survey on chatbots for PLWD, we found that these systems have generally limited scope based on the information programmed into them by their creators [6]. They also have limited natural language–understanding capabilities. A new promising chatbot platform, ChatGPT, which was introduced in late 2022, is trained on huge amounts of data and has superior natural language–understanding technology (see the *Review of Query Tools* section for details). ChatGPT has been shown to be able to answer complex questions on a wide range of topics and even change the style of the responses based on the user (eg, change the literacy level, make the answer funny, etc).

PLWD have a wide range of needs for which they rely on the web for answers. A seminal paper on web searching identified 3 types of search needs of web users: informational, transactional, and navigational [7]. Informational queries aim to acquire some information assumed to be present on 1 or more web pages (eg, “Is there a cure for dementia?”). Transactional queries try to perform an activity (eg, “Find good home care in Riverside, California.”). In navigational queries, users look for the web address of an organization (eg, “WebMD”). We did not consider navigational queries in our study as their answers are usually trivial and they are not common in conversational settings.

AI chatbots, such as ChatGPT, have accelerated the transition from keyword queries to question answering (QA) [8], where the goal is not to return a list of relevant pages but to answer the user’s question. Web search engines, such as Google, have also been slowly moving toward this direction. In particular, for some queries, the results page contains a short text snippet at the top of the page (also called answer box or quick answer or direct answer), as shown in Figure 1.

Figure 1. The answer displayed at the top of the results page on the Google search engine.



Is there a cure for dementia?

About 131,000,000 results (0.43 seconds)

There is currently no "cure" for dementia. In fact, because dementia is caused by different diseases it is unlikely that there will be a single cure for dementia. Research is aimed at finding cures for dementia-causing diseases, such as Alzheimer's disease, frontotemporal dementia and dementia with Lewy bodies.

NHS
https://www.nhs.uk › Health A to Z › Dementia guide

Is there a cure for dementia? - NHS

People also ask

Can a person recover from dementia?

How long can you live with dementia?

What is the best treatment for dementia?

What are main causes of dementia?

NHS
https://www.nhs.uk › Health A to Z › Dementia guide

What are the treatments for dementia? - NHS

There is currently no cure for dementia. But there are medicines and other treatments that can help with dementia symptoms.

Medicines To Treat Dementia · Acetylcholinesterase... · Medicines To Treat Related...

Review of Query Tools

Web search engines, such as Google and Bing, continuously collect (crawl) content (documents) from the web, store it locally, and index it for faster retrieval at query time. When a user submits a query, the search engine uses a ranking algorithm to assign a score to each document and generate a ranked list of results [9]. Recently, search engines have been trying to move from *keyword search*, where a user provides keywords and the search engine returns a list of pages, to *QA* [8], where the user asks a question (eg, “What is the most popular drug for depression?”) and the search engine returns the answer (eg, “selective serotonin reuptake inhibitors [SSRIs]”) on the top of the results page, followed by a classic list of pages.

The exact ranking formula used by web search engines is not public, but the general technologies used are known. Specifically, before 2018, ranking used a combination of numerous features, such as the number of times query keywords appear in a document, the incoming hyperlinks of the document, and the importance of the domain. These features are combined using a function learned through machine learning [9].

A breakthrough came in 2018, when Google published a paper on bidirectional encoder representations from transformers (BERT) [10], which is a deep learning model for text. Specifically, BERT creates a semantic model of a text segment (a word, sentence, or paragraph), which then facilitates a semantic comparison of this text to a query. Some of the key advantages of BERT over previous keyword-based methods are as follows: (1) Synonyms are considered, or more generally alternative ways to express the same meaning; (2) all the words in the text are considered instead of just focusing on the most important words, such as nouns; (3) the order of the words becomes important; and (4) the exact segment of a document that answers the query is returned instead of returning the whole document. BERT has significantly improved the quality of search results and has brought web search engines closer to the goal of QA.

ChatGPT is a conversational agent, or chatbot, that can answer user questions in a way similar to how a human would respond [11]. ChatGPT is built on top of the third-generation Generative Pre-trained Transformer (GPT-3), which is a *language model* that has been trained using about 500 billion tokens (terms or words) [12]. Most of these tokens come from web pages (a subset of the pages crawled by a web search engine, such as Google), Wikipedia, and books. Importantly, language models are expensive to train, and hence, they are trained infrequently (eg, once a year), which means they may not contain the most current information available.

A language model can be viewed as a tool to generate reasonable continuations for a text segment. For example, for the text "it is sunny in," a language model may suggest continuing this text with "California." The key improvement of ChatGPT over GPT-3 is that it tries to make the chatbot responses useful to the user; that is, the response should answer the user's question instead of just responding with some text that would naturally follow the user's text in a document. To achieve that, ChatGPT is trained (more accurately, fine-tuned in addition to the training that GPT-3 already has) using human feedback to become more useful than GPT-3 [13]. Specifically, ChatGPT is fine-tuned using a technology called *reinforcement learning from human feedback*, whereby it can refine its subsequent responses based on the perceived usefulness of the answer to the current user.

ChatGPT only periodically gets retrained, every several months or even years, because of the high cost of training; that is, the response of ChatGPT is not up to date if it refers to recent events.

Access to ChatGPT is available on the ChatGPT website [14]. New users need to sign up using their email. They are provided with a small number of credits, and then they must pay a per usage fee computed based on the number of tokens submitted to ChatGPT.

Potential of ChatGPT for Health Literacy

Although Google has been available for health information seeking for more than 2 decades, ChatGPT is a new resource that has the potential for promoting health literacy. However, given its recent availability to the public, there is limited research on how this technology may have practical applications in health care, especially for patient or caregiver education. Recently, Lee et al [15] suggested that ChatGPT has the potential to meet the needs of medical professionals related to medical note taking, answering medical problems for patient cases, and medical consultation [15]. In fact, the authors noted that when given a battery of medical test questions, ChatGPT provided correct responses 90% of the time. However, the information needs of patient and caregiver populations are quite different than those of trained medical professionals. For example, ChatGPT is not programmed to generate images, such as diagrams or other graphic tools, that a layperson may find helpful in understanding their health condition. However, an added feature of ChatGPT is that it allows users to easily change the style of a response to better match their profiles [13]. For example, one may ask, "Explain dementia to a 10-year-old child" or "Explain COVID-19 in a funny way" or "How would Obama explain a health care deductible?" To answer such questions, ChatGPT uses deep learning to paraphrase the responses, given the conversational model of the requested style.

There is no previous study analyzing the usefulness of ChatGPT for PLWD. This paper studied the potential of ChatGPT for health information seeking by PLWD. We investigated how ChatGPT compares to a web search for various types of query needs and presented our findings related to the strengths and weaknesses of each technology. We considered criteria such as reliability, accuracy, readability, and objectivity, which are critical for PLWD in evaluating health-related information.

Methods

Study Design

We developed a systematic strategy for data collection and analysis. This included identifying questions that caregivers commonly ask about Alzheimer disease and related dementias (ADRD) and caregiving, as well as establishing a set of criteria to assess the collected responses from ChatGPT and Google. More details on this process are outlined later.

Question Identification

We established 2 categories of questions that caregivers commonly ask, informational and transactional questions (ie, questions about service delivery), which followed a common classification, as discussed earlier [7]. We considered 2 of the 3 categories of web questions for the reasons discussed in the *Introduction* section.

For general information questions, we modified individual items from the Alzheimer's Disease Knowledge Scale (ADKS) [16]. The ADKS is a validated assessment tool that includes 30 statements about Alzheimer disease, and the participant selects *true* or *false* for each item. The statements address a number of topics related to assessment and diagnosis, caregiving, course, life impact, treatment and management, and symptoms. For example, consider the following 2 statements:

> *A person with Alzheimer disease becomes increasingly likely to fall down as the disease gets worse.*
>
> *Having high blood pressure may increase a person's risk of developing Alzheimer disease.*

To convert these statements to questions, we added a prefix of "Is it true that" to each—for example, "Is it true that a person with Alzheimer disease becomes increasingly likely to fall down as the disease gets worse?" The ADKS is 1 of the most used tools to assess overall Alzheimer disease–related knowledge. These knowledge items across multiple domains were developed to identify gaps in patient and caregiver knowledge, specifically for those with dementia-related concerns seeking a dementia evaluation [16], making the tool appropriate for those searching for information.

The 30 transactional items used in this study represent commonly used dementia-related services and support. Families and caregivers confronted with caring for a loved one with dementia will often need to seek information about how to identify trustworthy, quality, local, and affordable services; these 30 items represent a spectrum of dementia-related services. Specifically, we considered questions related to how PLWD look for services they need within their local community [17]. We narrowed the list to 6 common categories of services: *adult day care*, *home health care*, *hospice care*, *respite care*, *memory clinics*, and *nonemergency medical transportation (NEMT)*. For each of these services, we created questions about *quality*, *accessibility*, and *affordability*. Specifically, we generated 5 questions for each service type. For example, here is a list of the 5 questions generated for *adult day care*:

- How do I pick the best adult day care? (Quality)
- How do I find adult day care in Riverside, California? (Accessibility)
- How do I pay for adult day care in California? (Affordability)
- How much does adult day care cost per day in Riverside, California? (Affordability)
- Does California license adult day care? (Quality)

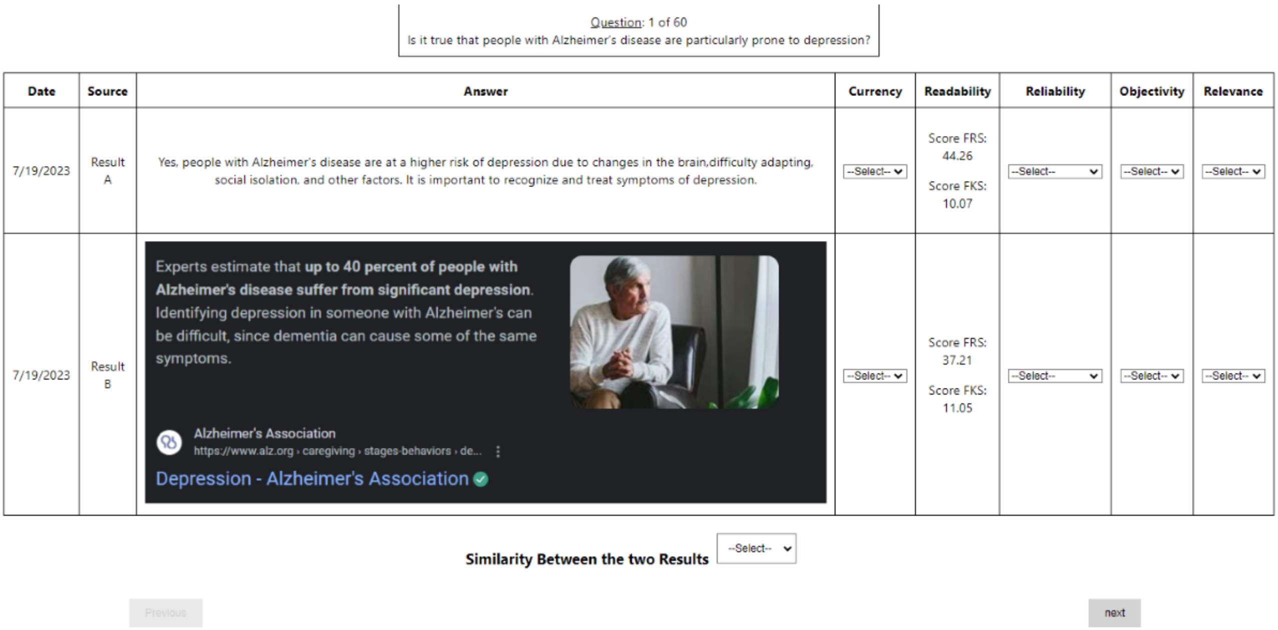
The same questions were generated for each of the remaining 5 service types.

We used Riverside, California, as the location for our queries. Hence, we had a total of 30 transactional questions. The list of all 60 questions is shown in Multimedia Appendix 1.

Data Collection and Management

One member of the research team (author SG) identified responses by entering each question into the ChatGPT and Google search engines between March 10 and 13, 2023, using a computer located in Riverside, California. Another team member (author VH) supervised SG and reviewed the collection process. All team members provided feedback on the process. Next, a database was created that displayed the questions, each response, and the assessment criteria. The displaying of ChatGPT responses is straightforward, as each response is in plain text (see Figure 2). For Google, we ignored advertisement results and showed the answer box generated (direct answer, as shown in Figure 2) if any existed, or else we showed the top organic (ie, nonadvertisement) result (title and snippet). Due to the different formatting (plain text for ChatGPT and rich text for Google), it was not possible to mask the source, but the raters were not told the name of the platforms and were asked to disregard this factor in their assessments.

Figure 2. Screenshot of the evaluation web application. Each rater selects a response from the drop-down menus and then advances to the next question.



Question: 1 of 60
Is it true that people with Alzheimer's disease are particularly prone to depression?

Date	Source	Answer	Currency	Readability	Reliability	Objectivity	Relevance
7/19/2023	Result A	Yes, people with Alzheimer's disease are at a higher risk of depression due to changes in the brain,difficulty adapting, social isolation, and other factors. It is important to recognize and treat symptoms of depression.	--Select--	Score FRS: 44.26 Score FKS: 10.07	--Select--	--Select--	--Select--
7/19/2023	Result B	Experts estimate that **up to 40 percent of people with Alzheimer's disease suffer from significant depression**. Identifying depression in someone with Alzheimer's can be difficult, since dementia can cause some of the same symptoms. Alzheimer's Association https://www.alz.org › caregiving › stages-behaviors › de... Depression - Alzheimer's Association	--Select--	Score FRS: 37.21 Score FKS: 11.05	--Select--	--Select--	--Select--

Similarity Between the two Results --Select--

Assessment Criteria

The rating system used was an extension of a rating system previously used to assess the content quality of dementia websites [18]. The 5 rating criteria we used are:

- Currency (yes, no, not sure): This criterion assessed whether information was provided on when the content was created and whether it was created within the past 5 years. This included the date of creation, the last updated date, and information in the actual response (eg, "In 2021, it was found that…"). Responses were deemed not current if they either lacked information about when the content was created or a date was provided but it was more than 5 years old. An assessment of "not sure" included cases where the response included text such as "Over the past year, it has been found that…," where information about the date of creation was provided but we could not assess when it was published.
- Readability: The information should be clear and easy to understand by users of different literacy levels. To assess readability, we computed the Flesch Reading Ease Score (FRS) and the Flesch–Kincaid Grade Level Score (FKS) for each response [19]. The FRS ranges from 1 to 100, where 100 is the highest level of readability. A score of 60 is considered standard for publications targeting a general audience, and a score of 70 or more is considered easy for the average adult to read [20]. The FKS is calculated by translating the FRS into a grade reading level. For example,

a score between 70 and 80 would translate to about an 8th-grade reading level.

- Reliability (reliable source, other source, no source, not sure): A response was deemed reliable if it provided the source of the content and the source was considered a reputable source for information about dementia, such as government websites and well-known organizations that provide information about dementia and caregiving (eg, Alzheimer Association). "Other source" referred to responses where a source was provided with the response (eg, website, organizational name) but it was not clear whether it was a reliable source. "No source" referred to responses that did not provide any information about the source. "Not sure" was used in cases where the response referred the reader to a reliable source for further information but there was no information about where the response content came from—for example, a response that read "Contact the Centers for Medicare and Medicaid for additional information" without offering information about the source of the response content.
- Objectivity (yes, no, not sure): A rating of "yes" reflected factual information that lacked any feelings or opinion about the content. "No" and "not sure" were assigned to cases where the content came from a for-profit agency (eg, Luxe Homecare) or if it was unclear whether the source was a for-profit organization.
- Relevance (high, medium, low, not sure): The raters evaluated the extent to which each response addressed the question asked. We also asked the raters to assess the similarity between the 2 results as "high," "medium," "low," or "not sure."

Assessment Process

Initially, a reviewer-training session was conducted with 3 of the authors (EB, NR, and SS). Differences in rater responses were discussed to ensure uniform application of assessment criteria [21]. During the training, the authors rated the first 15 questions together, that is, they jointly assigned a single rating. The remaining 45 questions were then rated independently by each rater.

Data Analysis

Frequencies and other descriptive statistics for all responses were generated. The Levene test was performed to assess equality in the variances between ChatGPT and Google readability scores on the FRS and FKS. This was followed by a 2-tailed Welch *t* test for independent samples to determine differences in the means of both groups. We computed the Fleiss κ to measure the agreement among the raters for the 45 questions that were rated independently.

Results

For reproducibility purposes, we saved the results of our queries on ChatGPT and Google using the 60 questions [22].

Currency

When comparing the responses as they related to the currency of information, we found that for ChatGPT, only 1 (1.7%) response listed the date when the data presented were collected, and the remaining 59 (98.3%) responses provided no dates. For Google, 19 (31.7%) responses provided dates when the data presented were collected or the dates when the source was updated, and dates fell within the past 5 years. The remaining 41 (68.3%) responses did not include dates when the information was provided or the dates provided were more than 5 years old. When comparing Google responses deemed current, there was no difference between informational (n=9, 30%) and transactional (n=10, 33.3%) questions. Table 1 illustrates the currency of both informational and transactional question responses.

Table 1. Currency for informational and transactional questions.

Currency	Transactional questions		Informational questions	
	ChatGPT, n (%)	Google, n (%)	ChatGPT, n (%)	Google, n (%)
Yes	0	10 (33.3)	0	9 (30.0)
No	30 (100.0)	20 (66.7)	30 (100.0)	21 (70.0)

Reliability of Information

Regarding reliability, ChatGPT did not identify any sources for the information it provided in its responses. However, 4 (6.7%) responses were rated as *not sure* because they did not cite any sources but directed the user to contact an agency deemed reputable. For example, for the question "How do I find home health care in Riverside, California?" ChatGPT instructed the reader to contact the Riverside County Office on Aging. In Google, all responses provided the websites from where the information was sourced. Among these sources, 36 (60%) were deemed reliable (eg, Alzheimer Association, National Council on Aging) and 24 (40%) were unfamiliar to the reviewers.

Table 2 depicts the results of our analysis of the reliability of both informational and transactional questions. Reliability, in this context, refers to the extent to which the information provided by the source can be trusted as accurate and credible. The table shows that there was little difference between the *reliable source* and *no source* responses for both transactional and informational questions. However, for informational questions, the *other source* and *not applicable* responses were noticeably higher compared to transactional questions.

Regarding currency and reliability, we asked ChatGPT a follow-up question ("Where did you get this information?"), and it is worth noting that in almost all cases, ChatGPT responded that it uses a wide range of sources to generate responses, with no further details.

Table 2. Reliability for informational and transactional questions.

Reliability	Transactional questions		Informational questions	
	ChatGPT, n (%)	Google, n (%)	ChatGPT, n (%)	Google, n (%)
No source	28 (93.3)	0	30 (100.0%)	0
Not sure	2 (6.7)	0	0	0
Reliable source	0	16 (53.3)	0	25 (83.3)
Other source	0	14 (46.7)	0	5 (16.7)

Objectivity

All the responses provided by ChatGPT were rated as objective, while 49 (81.7%) of the Google responses were deemed objective. When analyzed further, ChatGPT and Google performed similarly with regard to objectivity for informational questions. However, for transactional questions, 2 (3.3%) responses by Google were assessed as not being objective, because the source was a for-profit organization that provided services for PLWD, and it was unclear whether 11 (18.3%) of the sources were for-profit service agencies. Although the responses did not directly advertise services of the organization, the actual or potential conflicts of interest posed by the sources raised questions about the extent to which the response would be fully objective—for example, if a response to the query "How do I pick the best home health care?" was published on a website for an agency that provides home health care services.

Table 3 demonstrates the objectivity of the 2 types of questions used in this study. For informative questions, the responses were often more congruent, but they varied for transactional questions.

Table 3. Objectivity for informational and transactional questions.

Objectivity	Transactional questions		Informational questions	
	ChatGPT, n (%)	Google, n (%)	ChatGPT, n (%)	Google, n (%)
Yes	30 (100.0)	24 (82.7)	30 (100.0)	30 (100.0)
No	0	1 (3.5)	0	0
Not sure	0	4 (13.8)	0	0

Relevance

For ChatGPT, reviewers rated 58 (96.7%) responses as being *highly relevant* to the question asked, 2 (3.3%) were rated as having *medium relevance*, and none of the responses were assessed as having *low relevance.* For Google, reviewers rated 36 (60%) responses as being *highly relevant* to the question, 4 (6.7%) were rated as having *medium relevance*, and 20 (33.3%) were assessed as having *low relevance.*Table 4 shows the results broken down by type of question. A comparison showed that the relevance of responses by ChatGPT was similar for both types of questions, whereas Google responses were more likely to be rated as having *medium* (n=5, 8.3%) or *low* (n=16, 53%) relevance for transactional questions. For transactional questions, the relevance of Google responses was typically ranked lower because Google drew upon sources that were referral services for dementia care or service providers themselves.

Table 4. Relevance for informational and transactional questions.

Relevance	Transactional questions		Informational questions	
	ChatGPT, n (%)	Google, n (%)	ChatGPT, n (%)	Google, n (%)
High	30 (100.0)	22 (73.3)	28 (93.3)	23 (76.7)
Low	0	7 (23.3)	2 (6.7)	1 (3.3)
Medium	0	1 (3.3)	0	6 (20.0)

Similarity

Overall, the similarity between the content of ChatGPT and Google responses was rated as *high* for 13 (21.7%) responses, *medium* for 16 (26.7%) responses, and *low* for 31 (51.6%) responses. Our study found that the search results for informational questions were more similar than those for transactional questions, as shown in Table 5. These findings suggest that different types of queries may require different search strategies.

Table 5. Similarity for informational and transactional questions.

Similarity	Transactional questions, n (%)	Informational questions, n (%)
High	9 (32.1)	10 (33.3)
Medium	7 (25.0)	8 (26.7)
Low	12 (42.9)	12 (40.0)

Readability

The readability of responses for both ChatGPT and Google varied widely, as displayed in Table 6. For ChatGPT, no responses had FRS ≥60 (indicating general reading audiences), while for Google, 12 (20%) responses had FRS ≥60. This was also the case for the FKS, where ChatGPT had 3 (5%) scores and Google had 21 (35%) scores at an 8th grade readability level or lower. However, there was greater variability in the readability scores of Google responses. The Levene test showed statistically significant differences in the variances between the ChatGPT and Google responses for the FRS ($F_{2,118}$=11.16, P=.001) and the FKS ($F_{2,118}$=9.89, P=.002). Therefore, a subsequent 2-tailed Welch t test for independent samples was performed, which also found statistically significant differences in the FRS (t_{100}=–3.26, P=.001) and the FKS (t_{100}=4.44, P<.001). Overall, the responses provided by Google had easier readability for general audiences, with a mean FKS of 9.86 (SD 3.47), which indicated that, on average, the Google responses were written at a 9th-grade reading level compared to ChatGPT, which had a mean FKS of 12.17 (SD 1.94), or a 12th-grade reading level.

The findings of this study are summarized in Table 7.

Table 6. Summary of FRS[a] and FKS[b] readability scores.

Score and variables	ChatGPT (N=60)	Google (N=60)
FRS		
Mean (SD)	33.5 (12.46)	43.36 (19.61)
Variance	157.82	390.93
Range	4.84-57.81	4.11-91.84
FKS		
Mean (SD)	12.17 (1.94)	9.86 (3.47)
Variance	3.75	12.05
Range	7.48-16.11	2.71-19.39

[a] FRS: Flesch Reading Ease Score.

[b] FKS: Flesch–Kincaid Grade Level Score.

Table 7. Summary of response assessment for ChatGPT and Google.

Assessment criterion	Superior performance	Comments
Currency	Google	No dates provided in ChatGPT results
Reliability	Google	No sources provided by ChatGPT
Objectivity	ChatGPT	Google more likely to generate responses from for-profit service agencies
Relevance	ChatGPT	Google's relevance lower for informational queries
Readability	Google	Both ChatGPT and Google requiring relatively high-grade-level preparation

Interrater Agreement

We calculated the Fleiss κ to determine whether there was agreement between the 3 raters in their assessments of currency, reliability, objectivity, and relevance. Overall, there was good reliability of the ratings of reliability (κ=0.732, 95% CI 0.655-0.808, P<0.001), currency (κ=0.731, 95% CI 0.614-0.848, P<.001), and relevance (κ=0.618, 95% CI 0.518-0.718, P<.001). In addition, there was fair agreement among the raters' judgment of objectivity (κ=0.218, 95% CI 0.115-0.320, P<.001). The reason is that raters may have considered different ways for judging objectivity, including whether the source website was from a for-profit company, the content of the result, or even its tone.

Discussion

Principal Findings

The primary finding of this study is that ChatGPT and Google have complementary strengths for people with cognitive decline or their caregivers (Table 7). Specifically, ChatGPT provided more relevant responses to the queries (Google did not have an answer box in 8 of the 30 informational and 12 of the 30 transactional questions) and greater objectivity, given that Google was more likely to generate responses from sources that may have potential or actual financial conflicts of interest. Google has slowly been moving from keyword search to QA (see the *Review of Query Tools* section), which means that an increasing number of queries return answer boxes over time. Google has superior currency, as Google crawls web pages continuously to collect the latest content. However, ChatGPT only periodically gets retrained, as mentioned in the *Review of Query Tools* section. Further, both suffer from a lack of time stamps indicating how current a result is. Reliability is also a stronger point for Google, as it shows the sources (URLs) of each result in contrast to ChatGPT.

Reliability is critical, given that health misinformation on the internet has been shown to be a significant challenge for PLWD [23]. In some cases, perceptions of misinformation and scams result in some PLWD altogether avoiding searching for information about health-related topics on the internet [24]. PLWD have also reported frustration in reading about emerging research findings related to ADRD on the internet, only to find later that those findings were inaccurate as newer research is reported [23]. The impact of health misinformation on the internet on the dementia community has most recently been documented during the COVID-19 pandemic, where caregivers reported that they refused vaccination for themselves or their care recipients based on fears from what they read on the internet [25].

In addition to the currency and reliability limitations, we also observed a few other shortcomings of ChatGPT. First, it has limited support for other *languages* as its model has not been trained to the same level as in English. Second, it often *crashes* in the middle of a conversation or is unavailable (we expect this to improve over time). Third, it requires a monthly *subscription* to use (after a number of free accounts were given out), which is slowly changing as it is becoming part of the Microsoft Bing platform. Fourth, it is not able to help PLWD complete any *transactions*, such as administering a questionnaire to evaluate cognitive function or to find a doctor in their area. Fifth, when it does not fully understand a question, it does not ask any *clarification* question (eg, “Did you mean home care services?”). Instead, it responds as best it can based on the last input.

In this study, in terms of readability, ChatGPT responses tended to be too high in reading difficulty for a general audience, with readability ratings at an average 12th-grade reading level compared to Google’s average of a 9th-grade reading level. However, many of the Google responses were also more difficult to read than what is recommended for health education materials. The raters in this study were health and social service professionals with advanced education and training. Hence, their understanding of the ChatGPT responses may be higher compared to many PLWD. Health literacy, and more specifically digital health literacy, has been an ongoing challenge in the United States and is a priority for *Healthy People 2030*, which provides 10-year, measurable public health objectives and tools to help track progress toward achieving them [26]. Moreover, there are significant disparities in health and digital health literacy [26], which further exacerbates health inequalities for underserved populations of PLWD. Digital literacy has been identified as an ongoing challenge for PLWD, which was made more evident during the COVID-19 pandemic, when many caregivers needed to access digital platforms for health and support services [27]. Although the need for technology use during the pandemic may have increased digital literacy among this population, findings on this have not yet been reported. Prior research has found that there is a strong relationship between digital health literacy and overall health literacy for caregivers, which suggests that improving digital health literacy can improve outcomes such as self-efficacy for this population [28].

Although this analysis focused on the readability level of the responses generated, it also highlighted that many caregivers may have limits regarding health and computer literacy that impede their search strategies, such as the use of keywords. For instance, although the ChatGPT responses were deemed highly relevant to the questions posed in this study, caregivers with similar information needs may not generate the same responses due to less efficient queries.

The website of ChatGPT lists a few more limitations [11], which we include here for completeness:

> *(i) ChatGPT sometimes writes plausible-sounding but incorrect or nonsensical answers. (ii) ChatGPT is sensitive to tweaks to the input phrasing or attempting the same prompt multiple times. (iii) The model is often excessively verbose and overuses certain phrases. (iv) While we’ve made efforts to make the model refuse inappropriate requests, it will sometimes respond to harmful instructions or exhibit biased behavior.*

Ethical Issues

There are a number of ethical issues to consider regarding the use of ChatGPT and Google when searching for health information on the internet. For ChatGPT, 1 concern is that since the platform is trained on existing content on the internet, it may provide users with information that is inaccurate or highly biased [29]. User privacy has also been raised as an issue regarding AI browsing tools in general, with concerns about how user interactions with such technologies may be tracked [29]. Such ethical concerns have led to calls for a code of ethics for ChatGPT and similar AI tools [30]. Similarly, ethical issues regarding the use of Google for searching for health information have also been raised. For example, Google has been criticized for publicizing inaccurate public health data [31], concerns about privacy of user input [32], and biased algorithms that inform the top results of any given search query [33]. Hence, more work is needed on addressing ethical concerns in digital health literacy in general.

Complementary Nature

Although ChatGPT can provide health-related information with high levels of accuracy [15], it has several drawbacks. For example, due to the lack of information about the sources of responses, users may need to cross-reference the responses by consulting with additional sources. This may increase the burden of information seeking [15]. Therefore, 1 potential result is that PLWD may find themselves using ChatGPT and Google in tandem rather than relying on one source or the other.

Future Predictions

We expect that chatbot platforms, such as ChatGPT, and web search engines, such as Google, will slowly start to converge in the next few years. Specifically, chatbots will support timely results, show sources (eg, web page addresses), and offer multimedia (ie, images and videos) user interfaces. However, web search engines will increasingly move away from lists of pages to answers as results. For example, the results page (Figure 1) will increasingly emphasize the top answer (answer box), which will become more accurate, and downplay the subsequent list of web pages. Web search engines will also get better at exploiting the search (or conversational) history of the user.

Related Work

Current State of Dementia-Focused Chatbots

Previously, we conducted a systematic review of commercially available chatbot apps that targeted PLWD [6]. Overall, we found that few chatbots focus on dementia and most are designed for engaging with PLWD, such as guided reminiscing. All but 1 chatbot were Alexa (Amazon) skill apps, which limits the target audience to those with Alexa-enabled technologies. However, research literature is increasingly focusing on this topic. For example, Varshini et al [34] recently reported the development of a *companion chatbot* with several safety and supportive features, such as communicating with family members about the location of PLWD and providing care recipients with reminders for memory support. Jiménez et al [35] presented a model for using Alexa for caregiver support when the care recipient presents problematic behaviors. However, the application of chatbot technologies for dementia-related health education is underdeveloped.

Recently, we completed a pilot test of an educational and supportive app for caregivers, called CareHeroes, that is multifunctional and includes an educational chatbot [36]. Other educational and supportive tools that are available on the app include links to vetted websites from trusted sources for dementia and caregiving, educational videos developed by the team and offered in English and Spanish, self-assessments for burden and depression that provide feedback based on the responses, and clinical assessments of care recipients. The chatbot is programmed to respond to common questions related to dementia and caregiving, and responses are based on content from the book *The Dementia Caregiver*, by Dr Marc E Agronin, a geriatric psychiatrist [37]. In the study, caregivers were asked to use the app for a period of 12 months and the authors tracked usage data from CareHeroes. Overall, they found that the educational chatbot is the most used educational feature on the app, with caregivers using it to gather information about depression and sleep problems experienced by the patients they cared for and about living wills. The findings suggest that chatbot technologies offer an opportunity to provide targeted education content to caregivers based on their individual informational needs. However, more research needs to be conducted to advance work in this area.

Searching the Web for Health Information

Caregivers have been demonstrated to be more active than noncaregivers in seeking information about health-related topics and often use the web to gather information from websites and social media [1]. However, caregivers' experiences with using the web to gather information about health is not universal. For instance, a recent study analyzing data from the Health Information National Trends Survey (HINTS) found that 42.7% of caregivers in the United States do not have broadband access and there are significant disparities in access to broadband internet among caregivers [38]. This may limit access to web-based tools, such as ChatGPT and Google, especially for caregivers who may already be underserved. It has also been reported that caregivers from underserved populations, especially those who are immigrants, have more difficulty in finding the information they are seeking from web searches and are less likely to trust the information that they find [1]. To support caregivers in the future in health-related web searching, policy and research should focus more on advancing infrastructure for high-speed internet access and increasing digital health literacy among caregivers.

ChatGPT in Health Care

Research on using ChatGPT in health care apps is still in its infancy, although research on its potential use is increasing. A recent systematic review found that a number of potential benefits of using ChatGPT in health care settings have been identified in the literature, including improving health care services and health literacy, supporting research, and educating the health care workforce [30]. However, the same review identified a number of concerns with ChatGPT in health care apps, including ethical concerns around potential bias, the risk of spreading misinformation, and the security of protected health information. The literature on ChatGPT in health care has mostly focused on providers, educators, and researchers. Little has been done to explore its use for patient education. In their study using ChatGPT to answer questions about prostate cancer, Zhu et al [39] found that although all the large language models (LLMs) they submitted their questions to provided more than 90% accuracy, ChatGPT had the highest accuracy rate. They also noted that the free version of ChatGPT performs better than the paid version. However, they did not compare the responses from any of the LLMs with those from more traditional search engines, such as Google, as we set out to do. In addition, their study was similar to ours, in that they generated questions to ask the chatbots rather than examining how real PLWD interact with these search tools. Therefore, more studies are needed to better understand how ChatGPT and similar models can be appropriately used to improve health literacy.

Study Limitations

Although there were numerous considerations made to increase the rigor of this study, there are some limitations. First, the study was conducted in the midst of evolving AI technology; therefore, these findings are snapshots in time. Second, the service-related queries focused on a specific urban area, so we anticipate that these findings would be similar in other locations. Third, although we tried to identify frequently asked questions, the queries did not consider slang or cultural idioms and all queries were in English. Finally, the raters could guess which results corresponded to Google and which to ChatGPT, based on their presentation, as shown in Figure 2.

Conclusion

This paper studied how Google and ChatGPT compare in answering queries related to dementia or other cognitive decline. In total, 60 informational and transactional queries were selected, and their results were rated by 3 experts based on several criteria. We found that both Google and ChatGPT have strengths and weaknesses. Google more often provides the source and date of the response, whereas ChatGPT has a higher response accuracy and objectivity. Their combination could potentially provide results of higher quality. That is, more research and new technologies are needed that will leverage the language understanding and precision of ChatGPT and combine it with the wide coverage and currency of Google.

Acknowledgments

All authors declared that they had insufficient or no funding to support open access publication of this manuscript, including from affiliated organizations or institutions, funding agencies, or other organizations. JMIR Publications provided APF support for the publication of this article.

Data Availability

The data sets generated and analyzed during this study are available in Multimedia Appendix 1 and in a repository [22].

Conflicts of Interest

VH is the founder of SmartBot360, which is a health care chatbot company. SmartBot360 was not discussed in this paper.

Multimedia Appendix 1

List of questions.
[DOCX File , 16 KB-Multimedia Appendix 1]

References

1. Bangerter LR, Griffin J, Harden K, Rutten LJ. Health information-seeking behaviors of family caregivers: analysis of the Health Information National Trends Survey. JMIR Aging 2019 Jan 14;2(1):e11237 [FREE Full text] [doi: 10.2196/11237] [Medline: 31518309]
2. Wolff J, Darer J, Larsen K. Family caregivers and consumer health information technology. J Gen Intern Med 2016 Jan;31(1):117-121 [FREE Full text] [doi: 10.1007/s11606-015-3494-0] [Medline: 26311198]
3. Abner E, Jicha G, Christian W, Schreurs B. Rural-urban differences in Alzheimer's disease and related disorders diagnostic prevalence in Kentucky and West Virginia. J Rural Health 2016 Jun;32(3):314-320 [FREE Full text] [doi: 10.1111/jrh.12155] [Medline: 26515331]
4. Kaufman A, Kosberg J, Leeper J, Tang M. Social support, caregiver burden, and life satisfaction in a sample of rural African American and White caregivers of older persons with dementia. J Gerontol Soc Work 2010 Apr;53(3):251-269 [FREE Full text] [doi: 10.1080/01634370903478989] [Medline: 20336572]
5. Lee K, Hoti K, Hughes JD, Emmerton L. Dr Google and the consumer: a qualitative study exploring the navigational needs and online health information-seeking behaviors of consumers with chronic health conditions. J Med Internet Res 2014 Dec 02;16(12):e262 [FREE Full text] [doi: 10.2196/jmir.3706] [Medline: 25470306]
6. Ruggiano N, Brown EL, Roberts L, Framil Suarez CV, Luo Y, Hao Z, et al. Chatbots to support people with dementia and their caregivers: systematic review of functions and quality. J Med Internet Res 2021 Jun 03;23(6):e25006 [FREE Full text] [doi: 10.2196/25006] [Medline: 34081019]
7. Broder A. A taxonomy of web search. SIGIR Forum 2002 Sep;36(2):3-10 [FREE Full text] [doi: 10.1145/792550.792552]
8. Karpukhin V, Barlas Oguz SM, Patrick L, Ledell W, Sergey E, Danqi C, et al. Dense passage retrieval for open-domain question answering. 2020 Presented at: 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP); November 8-12, 2020; Punta Cana, Dominican Republic p. 6769-6781 [doi: 10.18653/v1/2020.emnlp-main.550]
9. Levene M. Search engines: information retrieval in practice. Comput J 2010 Apr 13;54(5):831-832 [doi: 10.1093/comjnl/bxq039]
10. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. 2019 Presented at: NAACL-HLT 2019; June 2-7, 2019; Minneapolis, MN p. 4171-4186
11. Introducing ChatGPT. OpenAI. 2022 Nov 30. URL: https://openai.com/blog/chatgpt/ [accessed 2023-05-11]
12. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language models are few-shot learners. 2020 Presented at: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020; December 6-12, 2020; Virtual p. 1877-1901

13. Ouyang L, Jeff W, Jiang X, Almeida D, Wainwright CL, Mishkin P, et al. Training language models to follow instructions with human feedback. arXiv:2203.02155 [cs.CL] posted online 2022 [doi: 10.48550/arXiv.2203.02155]
14. Welcome to ChatGPT. ChatGPT. URL: https://chat.openai.com/ [accessed 2023-07-17]
15. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. N Engl J Med 2023 Mar 30;388(13):1233-1239 [doi: 10.1056/nejmsr2214184]
16. Carpenter BD, Balsis S, Otilingam PG, Hanson PK, Gatz M. The Alzheimer's Disease Knowledge Scale: development and psychometric properties. Gerontologist 2009 Apr;49(2):236-247 [FREE Full text] [doi: 10.1093/geront/gnp023] [Medline: 19363018]
17. Finding dementia care and local services. U.S. National Institute on Aging. URL: https://www.alzheimers.gov/life-with-dementia/find-local-services [accessed 2023-04-29]
18. Zhu Y, Song T, Yu P. Developing methods to evaluate content quality of dementia websites. Stud Health Technol Inform 2021 Dec 15;284:374-378 [doi: 10.3233/SHTI210750] [Medline: 34920551]
19. Flesch RF. How to Write Plain English. New York, NY: Barnes & Noble; 1981.
20. Flesch R. A new readability yardstick. J Appl Psychol 1948 Jun;32(3):221-233 [FREE Full text] [doi: 10.1037/h0057532] [Medline: 18867058]
21. Waltz CF, Strickland OL, Lentz ER. Measurement in Nursing and Health Research. 4th ed. New York, NY: Springer Publishing; 2010.
22. Ganta SRR, Hristidis V, Brown EL, Ruggiano N. Memory query results evaluator, preview of the questions. Google Docs. 2023. URL: https://docs.google.com/spreadsheets/d/1fTbyJ3il0DRLlHGsSEal1uXCS9jJh7PI9gKc6WXxGqo/edit?usp=sharing [accessed 2023-07-17]
23. Dixon E, Anderson J, Blackwelder D, Radnofsky M, Lazar A. Barriers to online dementia information and mitigation. In: Proc SIGCHI Conf Hum Factor Comput Syst. 2022 Apr Presented at: CHI '22: CHI Conference on Human Factors in Computing Systems; April 29-May 5, 2022; New Orleans, LA [doi: 10.1145/3491102.3517554]
24. Dixon E, Piper AM, Lazar A. "Taking care of myself as long as I can": how people with dementia configure self-management systems. 2021 May Presented at: CHI '21: CHI Conference on Human Factors in Computing Systems; May 8-13, 2021; Yokohama, Japan [doi: 10.1145/3411764.3445225]
25. Savla J, Roberto KA, McCann BR, Blieszner R. COVID-19 vaccination experiences of family caregivers of persons living with dementia in rural Appalachia. J Appl Gerontol 2023 May;42(5):821-831 [FREE Full text] [doi: 10.1177/07334648221147916] [Medline: 36565159]
26. Jackson D, Trivedi N, Baur C. Re-prioritizing digital health and health literacy in Healthy People 2030 to affect health equity. Health Commun 2021 Sep;36(10):1155-1162 [FREE Full text] [doi: 10.1080/10410236.2020.1748828] [Medline: 32354233]
27. Chirico I, Giebel C, Lion K, Mackowiak M, Chattat R, Cations M, et al. Use of technology by people with dementia and informal carers during COVID-19: a cross-country comparison. Int J Geriatr Psychiatry 2022 Sep;37(9) [doi: 10.1002/gps.5801] [Medline: 36005276]
28. Efthymiou A, Middleton N, Charalambous A, Papastavrou E. Health literacy and eHealth literacy and their association with other caring concepts among carers of people with dementia: a descriptive correlational study. Health Soc Care Community 2022 May 06;30(3):1109-1119 [doi: 10.1111/hsc.13341] [Medline: 33956368]
29. Ray P. ChatGPT: a comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. IoT Cyber-Phys Syst 2023;3:121-154 [FREE Full text] [doi: 10.1016/j.iotcps.2023.04.003]
30. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. Healthcare (Basel) 2023 Mar 19;11(6):887 [FREE Full text] [doi: 10.3390/healthcare11060887] [Medline: 36981544]
31. Morley J, Machado CCV, Burr C, Cowls J, Joshi I, Taddeo M, et al. The ethics of AI in health care: a mapping review. Soc Sci Med 2020 Sep;260:113172 [doi: 10.1016/j.socscimed.2020.113172] [Medline: 32702587]
32. Esteve A. The business of personal data: Google, Facebook, and privacy issues in the EU and the USA. Int Data Privacy Law 2017:36-47 [FREE Full text] [doi: 10.1093/idpl/ipw026]
33. Lewandowski D, Sünkler S. What does Google recommend when you want to compare insurance offerings? A study investigating source distribution in Google’s top search results. AJIM 2019 May 31;71(3):310-324 [FREE Full text] [doi: 10.1108/ajim-07-2018-0172]
34. Varshini M, Surabhi S, Keerthan KT. The companion chatbot for dementia patients. Int J Adv Sci Tech 2020;29(4):6582-6592
35. Jiménez S, Favela J, Quezada, Alanis A, Castillo E, Villegas E. Alexa to support patients with dementia and family caregivers in challenging behaviors. In: Rocha A, Adeli H, Dzemyda G, Moreira F, editors. Information Systems and Technologies" WorldCIST 2022, Volume 1. Cham: Springer; 2022.
36. Ruggiano N, Brown EL, Clarke PJ, Roberts L, Daquin J, Agronin M, et al. Examining the clinical workflow and outcomes of integrating health IT to educate and support dementia caregivers. Agency for Healthcare Research and Quality. 2023 Mar. URL: https://digital.ahrq.gov/sites/default/files/docs/citation/r21hs026571-ruggiano-final-report-2022.pdf [accessed 2023-07-17]

37. Agronin M. The Dementia Caregiver: A Guide to Caring for Someone with Alzheimer's Disease and Other Neurocognitive Disorders. Lanham, MD: Rowman & Littlefield; 2015.
38. Kim H, Mahmood A, Goldsmith J, Chang H, Kedia S, Chang C. Access to broadband internet and its utilization for health information seeking and health communication among informal caregivers. J Med Syst 2021 Jan 15;45(2):24-29 [FREE Full text] [doi: 10.1007/s10916-021-01708-9] [Medline: 33452625]
39. Zhu L, Mou W, Chen R. Can the ChatGPT and other large language models with internet-connected database solve the questions and concerns of patient with prostate cancer and help democratize medical knowledge? J Transl Med 2023 Apr 19;21(1):269-264 [FREE Full text] [doi: 10.1186/s12967-023-04123-5] [Medline: 37076876]

Abbreviations

AI: artificial intelligence
ADKS: Alzheimer’s Disease Knowledge Scale
ADRD: Alzheimer disease and related dementias
BERT: bidirectional encoder representations from transformers
FKS: Flesch–Kincaid Grade Level Score
FRS: Flesch Reading Ease Score
GPT-3: third-generation Generative Pre-trained Transformer
LLM: large language model
PLWD: people living with dementia or other cognitive decline and their caregivers
QA: question answering

Edited by G Eysenbach, T de Azevedo Cardoso; submitted 15.05.23; peer-reviewed by A Garnett, T Hou; comments to author 02.06.23; revised version received 09.06.23; accepted 10.07.23; published 25.07.23

Please cite as:
Hristidis V, Ruggiano N, Brown EL, Ganta SRR, Stewart S
ChatGPT vs Google for Queries Related to Dementia and Other Cognitive Decline: Comparison of Results
J Med Internet Res 2023;25:e48966
URL: https://www.jmir.org/2023/1/e48966
doi: 10.2196/48966
PMID: