RESEARCH Open Access

Performance of large language models (LLMs) in providing prostate cancer information

Ahmed Alasker[1,2,3], Seham Alsalamah[2,3*], Nada Alshathri[2,3], Nura Almansour[2,3], Faris Alsalamah[2,3], Mohammad Alghafees[1,2,3], Mohammad AlKhamees[1,2,3,4] and Bader Alsaikhan[1,2,3]

Abstract

Purpose The diagnosis and management of prostate cancer (PCa), the second most common cancer in men worldwide, are highly complex. Hence, patients often seek knowledge through additional resources, including AI chatbots such as ChatGPT and Google Bard. This study aimed to evaluate the performance of LLMs in providing education on PCa.

Methods Common patient questions about PCa were collected from reliable educational websites and evaluated for accuracy, comprehensiveness, readability, and stability by two independent board-certified urologists, with a third resolving discrepancy. Accuracy was measured on a 3-point scale, comprehensiveness was measured on a 5-point Likert scale, and readability was measured using the Flesch Reading Ease (FRE) score and Flesch–Kincaid FK Grade Level.

Results A total of 52 questions on general knowledge, diagnosis, treatment, and prevention of PCa were provided to three LLMs. Although there was no significant difference in the overall accuracy of LLMs, ChatGPT-3.5 demonstrated superiority over the other LLMs in terms of general knowledge of PCa ($p = 0.018$). ChatGPT-4 achieved greater overall comprehensiveness than ChatGPT-3.5 and Bard ($p = 0.028$). For readability, Bard generated simpler sentences with the highest FRE score (54.7, $p < 0.001$) and lowest FK reading level (10.2, $p < 0.001$).

Conclusion ChatGPT-3.5, ChatGPT-4 and Bard generate accurate, comprehensive, and easily readable PCa material. These AI models might not replace healthcare professionals but can assist in patient education and guidance.

Keywords Prostate cancer, Artificial intelligence, Large language models, Chatbot, ChatGPT

Introduction

Large language models (LLMs) are natural language processing models that utilize deep learning algorithms to generate and present text in a human-like fashion. The Generative Pretrained Transformers (ChatGPT) model is a recent large-language artificial intelligence (AI) model. [1] Although ChatGPT was only recently introduced at the end of 2022, it has attracted much interest. ChatGPTs can carry out a wider range of natural language tasks than can prior deep learning AI models. In addition, it can generate chatty responses to user input that resemble human responses based on a wealth of data. [2]

*Correspondence:
Seham Alsalamah
seham1alslamh@gmail.com
[1] Division of Urology, Department of Surgery, Ministry of National Guard - Health Affairs, Riyadh, Saudi Arabia
[2] King Abdullah International Medical Research Center (KAIMRC), Riyadh, Saudi Arabia
[3] College of Medicine, King Saud bin Abdulaziz University for Health Sciences, Riyadh, Saudi Arabia
[4] Department of Surgical Specialities, College of Medicine, Majmaah University, Majmaah, Saudi Arabia

Therefore, ChatGPT has the potential to help people and communities make educated decisions about their health. [3] Nonetheless, ChatGPT has shown imperfections in providing medical answers, mainly due to the outdated data from September 2021 and before. [4] The current excitement and enthusiasm surrounding AI large language model chatbots drove Google to experiment with conversational AI through the Bard chatbot, released in 2023. It is powered by the Language Model for Dialogue Applications (LaMDA), invented by Google in 2017.

PCa is the second most common cancer in men worldwide, with an estimated prevalence of 43% in Saudi Arabia. [5, 6] PCa patients might present with localized symptoms or advanced disease. The diagnosis of PCa relies on digital rectal examination (DRE), prostate-specific antigen (PSA) analysis, and prostate biopsy. Management options for PCa include active surveillance, radiation therapy, and radical prostatectomy. Patients with more severe diseases, such as relapses or metastases, might require androgen deprivation therapy (ADT), salvage radiotherapy, and chemotherapy. [7] Due to the complexity of PCa diagnosis and management, patients often seek knowledge through additional resources such as AI chatbots; therefore, the ability of these LLMs to provide accurate, sufficient, and comprehensible information on PCa must be evaluated.

Methods

Common questions on PCa were collected from reliable websites that provide educational material to the general public, such as the American Society of Clinical Oncology (ASCO) or Prostate Cancer UK, Centers for Disease Control and Prevention (CDC), and Prostate Cancer Foundation (PCF). The selection criteria for questions were that the questions (1) target general knowledge (i.e. signs, symptoms, and pathophysiology), diagnosis, treatment, or prevention material on PCa and (2) be frequently asked by patients and the public as evaluated by board-certified urologists. The questions were then provided to three LLMs (ChatGPT-3.5, ChatGPT-4, and Google Bard) which were chosen due to their availability and accessibility. The factors used to assess the quality of responses were accuracy, comprehensiveness, readability, and stability. All responses were generated and recorded on 31/July/2023. To generate the text, we used ChatGPT-3.5, ChatGPT-4, and Google Bard, available at https://chat.openai.com/chat and https://bard.google.com/chat.

A 3-point scale was used for accuracy: one represents correct, two represents mixed with correct and incorrect/outdated data, and three represents completely incorrect data. A 5-point Likert scale was used for comprehensiveness of the responses, with one for "very comprehensive" and five for "very inadequate". For readability, the output answers were analysed for their sentences, words, syllables per word, and words per sentence. Moreover, the Flesch Reading Ease score and Flesch–Kincaid Grade Level were calculated for each text using the online calculator available at https://charactercalculator.com/flesch-reading-ease/ website. A higher Flesch Reading Ease score indicates an easily readable text, while the Flesch–Kincaid Grade Level indicates the grade-school level necessary to understand the text. [8] Due to the variety of responses generated for the same question by the LLMs, the stability of the output text was assessed for a select number of questions. Stability was determined based on the subjective assessment of whether the second and third answers were accurate compared to the first generated answer by two independent reviewers. Three responses were generated for 30 questions, and the chat history was read after each trial. Two experienced board-certified urologists worked independently to complete the ratings according to the National Comprehensive Cancer Network (NCCN), American Urological Association (AUA), and European Association of Urology(EAUAU) guidelines. [9–11] Discrepancies in grading and assessment among the two reviewers were independently reviewed and resolved by a blinded third board-certified urologist.

Statistical analysis

Statistical analysis was carried out using RStudio (version 4.3.0). We expressed categorical variables, including accuracy, comprehensiveness, readability, and stability, as frequencies and percentages. The significant differences between LLMs for those variables were assessed using Pearson's chi-square test or Fisher's exact test. We used the median and interquartile range (IQR) to present numerical variables, including words, sentences, syllables, words/sentences, syllables/words, FRE scores, and FK reading levels. The Kruskal−Wallis test was applied to explore the significant differences between the three LLMs in terms of the numerical variables. $p<0.05$ indicated statistical significance.

Results

A total of 52 questions were provided to three LLMs (ChatGPT-3.5, ChatGPT-4 and Google Bard). Most of the questions were acquired from ASCO (53.8%), the CDC (9.6%), Prostate Cancer UK (32.7%), and the PCF (3.8%). For each LLM, nine questions related to general knowledge (17.3%), five questions about diagnosis (9.6%), 27 questions about treatment (51.9%), and 11 questions about screening and prevention (21.2%).

Analysis of the accuracy of different LLMs

ChatGPT-3.5 achieved correct responses in 82.7% of cases, ChatGPT-4 in 78.8%, and Google Bard in 63.5%,

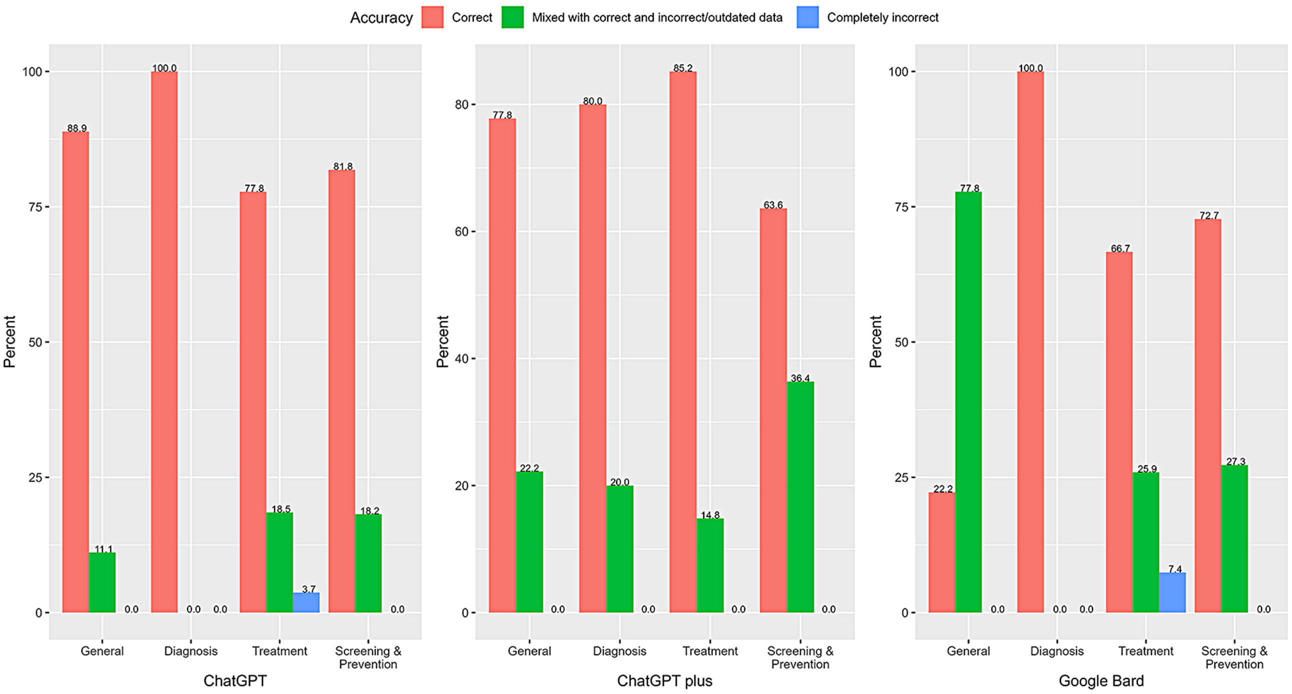

Fig. 1 The percentages of correct answers provided by each LLM

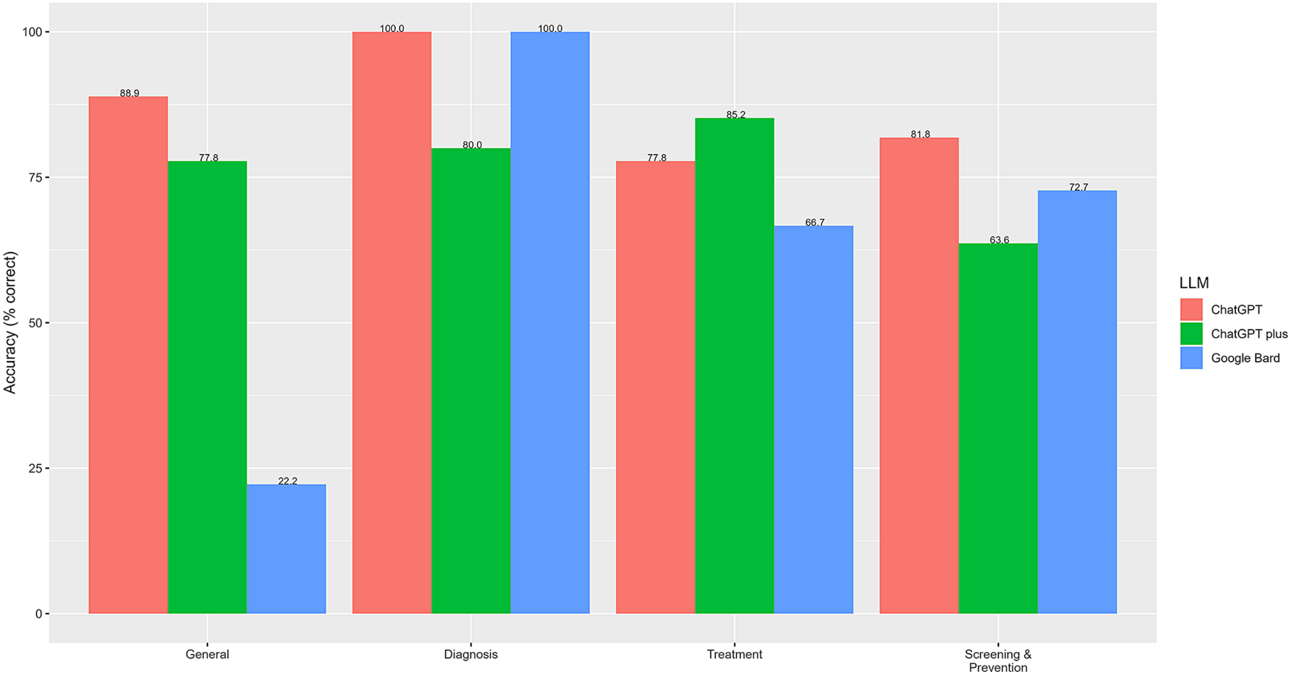

Fig. 2 Analysis of the accuracy of each LLM

with no significant difference in overall accuracy between LLMs (p=0.100). In the context of general knowledge questions, there was a statistically significant difference in accuracy among the LLMs (p=0.018; Fig. 1). ChatGPT-3.5 correctly answered 88.9% of the queries, ChatGPT-4 77.8%, and Google Bard 22.2% (Fig. 2). The accuracy of the diagnosis-related responses was not significantly different (p>0.999), with 100% for ChatGPT-3.5 and Google Bard and 80% for ChatGPT-4. For treatment-related questions, there were no significant differences in accuracy (p=0.496), with ChatGPT-3.5 achieving 77.8% accuracy, ChatGPT-4 85.2%, and Google

Bard 66.7%. Similarly, in the screening and prevention category, there were no significant differences in accuracy ($p=0.884$), with a score of 81.8% for ChatGPT-3.5, 63.6% for ChatGPT-4, and 72.7% for Google Bard (Table 1).

Analysis of the comprehensiveness of different LLMs

The overall comprehensiveness of the LLMs displayed statistically significant differences ($p=0.028$). Specifically, ChatGPT-4 achieved a significantly greater proportion of comprehensive responses (67.3%) than did ChatGPT-3.5 (40.4%) and Google Bard (48.1%). However, no significant differences were noted in the comprehensiveness of LLMs based on questions related to general knowledge, diagnosis, treatment, or screening and prevention (Table 2).

Analysis of the readability of different LLMs

The overall grade-level analysis revealed statistically significant differences among the LLMs ($p<0.001$). Specifically, Google Bard displayed a significantly greater percentage of responses rated at the 10th to 12th grade (34.6%) than did ChatGPT-3.5 (11.8%) and ChatGPT-4 (17.3%). Conversely, ChatGPT-4 demonstrated a significantly greater percentage of responses rated at the college level (61.5%) than did the Google Bard (36.5%). In the context of general knowledge about PCa, ChatGPT-4 exhibited more college-level responses (55.6%) than did Google Bards (0.0%); however, the difference was not statistically significant ($p=0.094$). For diagnosis-related questions, the analysis yielded a significant difference ($p=0.033$), with Google Bard producing a greater proportion of 10th- to 12th-grade responses (60.0%) than ChatGPT-4 (20.0%) and ChatGPT-3.5 (0.0%). In the treatment category, significant differences were observed ($p<0.001$), with ChatGPT-4 achieving a greater proportion of college-level responses (70.4%) than ChatGPT-3.5 (48.1%) and Google Bard (48.1%). Additionally, ChatGPT-3.5 had more college graduate-level responses (44.4%) than ChatGPT-4 (29.6%) and Google Bards (3.7%). In the context of screening and prevention, the difference between LLMs was not statistically significant (Table 3).

For the reading note, the analysis revealed statistically significant differences among the LLMs ($p<0.001$). Specifically, Google Bard displayed a significantly lower proportion of responses categorized as “Difficult to read” (36.5%) than did ChatGPT-3.5 (51.0%) and ChatGPT-4 (61.5%). In the “Very difficult to read” category, a significantly higher proportion (33.3%) compared to Google Bard (1.9%) and ChatGPT-4 (19.2%). In the diagnosis context, a significant difference was observed ($p=0.044$), with ChatGPT-3.5 producing a greater proportion of “Difficult to read” responses (75.0%) than ChatGPT-4 (60.0%) and Google Bard (0.0%). In the treatment category, significant differences were observed ($p<0.001$), with ChatGPT-4 achieving a greater proportion of “Difficult to read” responses (70.4%) than ChatGPT-3.5 (48.1%) and Google Bard (48.1%). There was no statistical significance in the screening and prevention context ($p=0.245$; Table 4).

Table 1 Accuracy of different LLMs

Characteristic	ChatGPT-3.5	ChatGPT-4	Google Bard	*p*-value
Overall ($n=52$)				0.100
Correct	43 (82.7%)	41 (78.8%)	33 (63.5%)	
Mixed	8 (15.4%)	11 (21.2%)	17 (32.7%)	
Completely incorrect	1 (1.9%)	0 (0.0%)	2 (3.8%)	
General ($n=9$)				**0.018**
Correct	8 (88.9%)	7 (77.8%)	2 (22.2%)	
Mixed	1 (11.1%)	2 (22.2%)	7 (77.8%)	
Completely incorrect	0 (0.0%)	0 (0.0%)	0 (0.0%)	
Diagnosis ($n=5$)				> 0.999
Correct	5 (100.0%)	4 (80.0%)	5 (100.0%)	
Mixed	0 (0.0%)	1 (20.0%)	0 (0.0%)	
Completely incorrect	0 (0.0%)	0 (0.0%)	0 (0.0%)	
Treatment ($n=27$)				0.496
Correct	21 (77.8%)	23 (85.2%)	18 (66.7%)	
Mixed	5 (18.5%)	4 (14.8%)	7 (25.9%)	
Completely incorrect	1 (3.7%)	0 (0.0%)	2 (7.4%)	
Screening & Prevention ($n=11$)				0.884
Correct	9 (81.8%)	7 (63.6%)	8 (72.7%)	
Mixed	2 (18.2%)	4 (36.4%)	3 (27.3%)	
Completely incorrect	0 (0.0%)	0 (0.0%)	0 (0.0%)	

Analysis of the accuracy of different LLMs in all categories, general knowledge, diagnosis, treatment, and screening and prevention

Table 2 Comprehensiveness of different LLMs

Characteristic	ChatGPT-3.5	ChatGPT-4	Google Bard	*p*-value
Overall (*n* = 52)				**0.028**
Very inadequate	0 (0.0%)	0 (0.0%)	2 (3.8%)	
Inadequate	19 (36.5%)	7 (13.5%)	13 (25.0%)	
Neither comprehensive nor inadequate	12 (23.1%)	8 (15.4%)	11 (21.2%)	
Comprehensive	21 (40.4%)	35 (67.3%)	25 (48.1%)	
Very comprehensive	0 (0.0%)	2 (3.8%)	1 (1.9%)	
General (*n* = 9)				0.520
Very inadequate	0 (0.0%)	0 (0.0%)	0 (0.0%)	
Inadequate	0 (0.0%)	0 (0.0%)	1 (11.1%)	
Neither comprehensive nor inadequate	1 (11.1%)	0 (0.0%)	1 (11.1%)	
Comprehensive	8 (88.9%)	7 (77.8%)	7 (77.8%)	
Very comprehensive	0 (0.0%)	2 (22.2%)	0 (0.0%)	
Diagnosis (*n* = 5)				0.301
Very inadequate	0 (0.0%)	0 (0.0%)	0 (0.0%)	
Inadequate	3 (60.0%)	1 (20.0%)	1 (20.0%)	
Neither comprehensive nor inadequate	1 (20.0%)	0 (0.0%)	0 (0.0%)	
Comprehensive	1 (20.0%)	4 (80.0%)	4 (80.0%)	
Very comprehensive	0 (0.0%)	0 (0.0%)	0 (0.0%)	
Treatment (*n* = 27)				0.064
Very inadequate	0 (0.0%)	0 (0.0%)	2 (7.4%)	
Inadequate	11 (40.7%)	5 (18.5%)	9 (33.3%)	
Neither comprehensive nor inadequate	8 (29.6%)	4 (14.8%)	5 (18.5%)	
Comprehensive	8 (29.6%)	18 (66.7%)	10 (37.0%)	
Very comprehensive	0 (0.0%)	0 (0.0%)	1 (3.7%)	
Screening & Prevention (*n* = 11)				0.331
Very inadequate	0 (0.0%)	0 (0.0%)	0 (0.0%)	
Inadequate	5 (45.5%)	1 (9.1%)	2 (18.2%)	
Neither comprehensive nor inadequate	2 (18.2%)	4 (36.4%)	5 (45.5%)	
Comprehensive	4 (36.4%)	6 (54.5%)	4 (36.4%)	
Very comprehensive	0 (0.0%)	0 (0.0%)	0 (0.0%)	

Analysis of the comprehensiveness of different LLMs in all categories, general knowledge, diagnosis, treatment, and screening and prevention

Notably, significant differences were observed among the LLMs for all the continuous parameters, including words, sentences, syllables, words/sentences, syllables/words, FRE scores, and FK reading levels ($p < 0.001$ for all; Table 5). First, when comparing the LLMs, ChatGPT-3.5 exhibited the fewest words (197.0), followed by Google Bard (290.0), while ChatGPT-4 had the most words (297.0). This trend suggested an increase in the number of words from ChatGPT-3.5 to ChatGPT-4 to Google Bard. Second, in terms of sentences, ChatGPT-3.5 had the lowest count (9.0), followed by ChatGPT-4 (15.5), and Google Bard had the highest (16.5). This indicates a gradual increase in the number of sentences from ChatGPT-3.5 to ChatGPT-4 to Google Bard.

Regarding syllables, ChatGPT-3.5 had the fewest (333.0), Google Bard had more (463.0), and ChatGPT-4 had the most (527.0), and This finding demonstrated a pattern of increasing syllables from ChatGPT-3.5 to Google Bard to ChatGPT-4. For the word/sentence ratio, ChatGPT-3.5 had the highest ratio (22.4), followed by ChatGPT-4 (19.2), and Google Bard having the lowest (18.3). Thus, the trend is a decrease in the word/sentence ratio from ChatGPT-3.5 to ChatGPT-4 to Google Bard. Similarly, for the syllable/word ratio, ChatGPT-3.5 had the highest ratio (1.8), followed by ChatGPT-4 (1.7) and Google Bard (1.6). Finally, in terms of readability, Google Bard had the highest FRE score (54.7), ChatGPT-4 had a midrange score (40.3), and ChatGPT-3.5 had the lowest (34.8). For the FK Reading Level, Google Bard had the lowest level (10.2), ChatGPT-4 had an intermediate level (12.3), and ChatGPT-3.5 had the highest level (14.0).

Analysis of the stability of different LLMs

The analysis of stability was exclusively performed on ten questions in each LLM. These included three inquiries related to diagnosis, three related to treatment, and four related to screening and prevention. Inconsistency was detected only in the response to one ChatGPT question about screening and prevention. There were no significant differences in the stability of LLMs in terms of any of the domains (Table 6).

Table 3 Grade levels of different LLMs

Characteristic	ChatGPT-3.5	ChatGPT-4	Google Bard	*p*-value
Overall (*n* = 52)				< 0.001
7th grade	0 (0.0%)	0 (0.0%)	2 (3.8%)	
8th & 9th grade	2 (3.9%)	1 (1.9%)	12 (23.1%)	
10th to 12th grade	6 (11.8%)	9 (17.3%)	18 (34.6%)	
College	26 (51.0%)	32 (61.5%)	19 (36.5%)	
College graduate	17 (33.3%)	10 (19.2%)	1 (1.9%)	
General (*n* = 9)				0.094
7th grade	0 (0.0%)	0 (0.0%)	2 (22.2%)	
8th & 9th grade	2 (22.2%)	1 (11.1%)	3 (33.3%)	
10th to 12th grade	3 (33.3%)	3 (33.3%)	4 (44.4%)	
College	2 (22.2%)	5 (55.6%)	0 (0.0%)	
College graduate	2 (22.2%)	0 (0.0%)	0 (0.0%)	
Diagnosis (*n* = 5)				0.033
7th grade	0 (0.0%)	0 (0.0%)	0 (0.0%)	
8th & 9th grade	0 (0.0%)	0 (0.0%)	2 (40.0%)	
10th to 12th grade	0 (0.0%)	1 (20.0%)	3 (60.0%)	
College	3 (75.0%)	3 (60.0%)	0 (0.0%)	
College graduate	1 (25.0%)	1 (20.0%)	0 (0.0%)	
Treatment (*n* = 27)				< 0.001
7th grade	0 (0.0%)	0 (0.0%)	0 (0.0%)	
8th & 9th grade	0 (0.0%)	0 (0.0%)	6 (22.2%)	
10th to 12th grade	2 (7.4%)	0 (0.0%)	7 (25.9%)	
College	13 (48.1%)	19 (70.4%)	13 (48.1%)	
College graduate	12 (44.4%)	8 (29.6%)	1 (3.7%)	
Screening & Prevention (*n* = 11)				0.235
7th grade	0 (0.0%)	0 (0.0%)	0 (0.0%)	
8th & 9th grade	0 (0.0%)	0 (0.0%)	1 (9.1%)	
10th to 12th grade	1 (9.1%)	5 (45.5%)	4 (36.4%)	
College	8 (72.7%)	5 (45.5%)	6 (54.5%)	
College graduate	2 (18.2%)	1 (9.1%)	0 (0.0%)	

Analysis of the grade-level scores of different LLMs in all categories: general knowledge, diagnosis, treatment, and screening and prevention

Discussion

This study aimed to compare the performance of three LLMs in response to PCa inquiries, and the results demonstrated interesting variability in terms of accuracy, comprehensiveness, readability, and stability. Although the evaluation of the overall accuracy of LLMs showed no significant difference, ChatGPT demonstrated superiority in most contexts. These findings align with previous studies that reached a similar conclusion, which showcases the capability of LLMs to provide accurate, but not optimal, answers to PCa patients. [12, 13] For the general knowledge questions, unlike Google Bard, which has poor accuracy, ChatGPT exhibited more remarkable performance, signifying its potential as a valuable tool that aids in patient education. Interestingly, in the context of treatment, all LLMs showed similar accuracy to that of ChatGPT-4 in the lead. The similar percentages between ChatGPT and Bard in the context of therapy could be due to the focused approach to these inquiries, which requires additional information without the need for inference. This finding aligns with that of a previous study that showed that Google Bard had inferior diagnostic skills to physicians since it requires excellent clinical reasoning and inferential abilities. [14] In regard to diagnosis, A study that analyzed the accuracy of ChatGPT's responses to PCa-related inquiries demonstrated that the worst performance was in the area of diagnosis alongside treatment. [15] However, our study showed that all LLMs had promising outcomes with no significant differences, which highlights the possibility of using LLMs in the context of formulating approaches to aid physicians in their diagnosis. In a study that compared ER physicians and ChatGPT in terms of diagnosing patients and triaging them, ChatGPT displayed accurate diagnoses in 87.5% of the cases, which further solidifies its applicability in this field. [16] Last, similar to the previous domain, the screening and prevention domain also demonstrated ChatGPT-4 pre-eminence with no significant overall differences among the three LLMs. These findings conciliate the general findings observed in this study, which is that ChatGPT is a superior model because of its ability to provide accurate responses.

Table 4 Analysis of the reading notes of different LLMs

Characteristic	ChatGPT-3.5	ChatGPT-4	Google Bard	*p*-value
Overall (*n* = 52)				< 0.001
Plain English	2 (3.9%)	1 (1.9%)	12 (23.1%)	
Fairly easy to read	0 (0.0%)	0 (0.0%)	2 (3.8%)	
Difficult to read	26 (51.0%)	32 (61.5%)	19 (36.5%)	
Fairly difficult to read	6 (11.8%)	9 (17.3%)	18 (34.6%)	
Very difficult to read	17 (33.3%)	10 (19.2%)	1 (1.9%)	
General (*n* = 9)				0.105
Plain English	2 (22.2%)	1 (11.1%)	3 (33.3%)	
Fairly easy to read	0 (0.0%)	0 (0.0%)	2 (22.2%)	
Difficult to read	2 (22.2%)	5 (55.6%)	0 (0.0%)	
Fairly difficult to read	3 (33.3%)	3 (33.3%)	4 (44.4%)	
Very difficult to read	2 (22.2%)	0 (0.0%)	0 (0.0%)	
Diagnosis (*n* = 5)				0.044
Plain English	0 (0.0%)	0 (0.0%)	2 (40.0%)	
Fairly easy to read	0 (0.0%)	0 (0.0%)	0 (0.0%)	
Difficult to read	3 (75.0%)	3 (60.0%)	0 (0.0%)	
Fairly difficult to read	0 (0.0%)	1 (20.0%)	3 (60.0%)	
Very difficult to read	1 (25.0%)	1 (20.0%)	0 (0.0%)	
Treatment (*n* = 27)				< 0.001
Plain English	0 (0.0%)	0 (0.0%)	6 (22.2%)	
Fairly easy to read	0 (0.0%)	0 (0.0%)	0 (0.0%)	
Difficult to read	13 (48.1%)	19 (70.4%)	13 (48.1%)	
Fairly difficult to read	2 (7.4%)	0 (0.0%)	7 (25.9%)	
Very difficult to read	12 (44.4%)	8 (29.6%)	1 (3.7%)	
Screening & Prevention (*n* = 11)				0.245
Plain English	0 (0.0%)	0 (0.0%)	1 (9.1%)	
Fairly easy to read	0 (0.0%)	0 (0.0%)	0 (0.0%)	
Difficult to read	8 (72.7%)	5 (45.5%)	6 (54.5%)	
Fairly difficult to read	1 (9.1%)	5 (45.5%)	4 (36.4%)	
Very difficult to read	2 (18.2%)	1 (9.1%)	0 (0.0%)	

Analysis of the reading notes of different LLMs in all categories, general knowledge, diagnosis, treatment, and screening and prevention

Table 5 Readability of LLMs

Characteristic	ChatGPT-3.5	ChatGPT-4	Google Bard	*p*-value
Words	197.0 (166.0–242.0)	297.0 (265.5–342.0)	290.0 (257.3–351.5)	< 0.001
Sentences	9.0 (7.0–11.0)	15.5 (13.0–18.3)	16.5 (13.0–20.3)	< 0.001
Syllables	333.0 (289.5–411.5)	527.0 (458.8–574.3)	463.0 (404.0–551.8)	< 0.001
Word/sentence	22.4 (20.4–24.7)	19.2 (17.5–20.7)	18.3 (16.0–20.0)	< 0.001
Syllable/word	1.8 (1.7–1.8)	1.7 (1.7–1.8)	1.6 (1.5–1.7)	< 0.001
FRE Score	34.8 (28.7–45.0)	40.3 (33.4–45.8)	54.7 (46.0–60.2)	< 0.001
FKGL	14.0 (12.2–15.2)	12.3 (11.3–14.0)	10.2 (9.1–11.6)	< 0.001

Analysis of the reading parameters and FRE and FKGL of different LLMs

Our study demonstrated a significant difference in overall comprehensiveness between ChatGPT-3.5, ChatGPT-4, and Google Bard. Lim et al. evaluated the performance of ChatGPT-3.5, ChatGPT-4, and Google Bard in generating comprehensive responses. They found no significant difference between the three LLM-Chatbots when comparing the comprehensiveness scores based on common queries answered by the three bots. [17] Our study proved that ChatGPT-4 had the highest number of comprehensive responses. On the other hand, Zhu et al. documented ChatGPT-3.5 as the LLM, which demonstrated the superior performance of providing the highest proportion of comprehensive responses, with 95.45% comprehensiveness. [12] As reported by Xie et al., who compared the comprehensibility of providing clinical guidance to junior doctors among three LLMs (including ChatGPT-4 and Google Bard), ChatGPT-4 performed best in generating comprehensive responses. [18] This finding aligns with our study, which proved that

Table 6 Stability of different LLMs

Characteristic	ChatGPT-3.5	ChatGPT-4	Google Bard	*p*-value
Overall (*n* = 10)				> 0.999
Consistent	9 (90.0%)	10 (100.0%)	10 (100.0%)	
Inconsistent	1 (10.0%)	0 (0.0%)	0 (0.0%)	
Diagnosis (*n* = 3)				
Consistent	3 (100.0%)	3 (100.0%)	3 (100.0%)	NA
Inconsistent	0 (0.0%)	0 (0.0%)	0 (0.0%)	
Treatment (*n* = 3)				
Consistent	3 (100.0%)	3 (100.0%)	3 (100.0%)	NA
Inconsistent	0 (0.0%)	0 (0.0%)	0 (0.0%)	
Screening & Prevention (*n* = 4)				> 0.999
Consistent	3 (75.0%)	4 (100.0%)	4 (100.0%)	
Inconsistent	1 (25.0%)	0 (0.0%)	0 (0.0%)	

Analysis of the stability of different LLMs in all categories, general knowledge, diagnosis, treatment, and screening and prevention

ChatGPT-4 was the highest-ranking LLM for generating comprehensive responses.

Google Bard provided more easily readable answers, achieved higher FRE and lower FKGL scores and generated adequate, straightforward sentences. These findings align with those of several studies illustrating a college level of ChatGPT answers. [19, 20] For instance, Cocci et al. analysed ChatGPT's responses to urology case studies and reported that ChatGPT achieved a college graduate reading level with median FRE and FKGL scores of 18 and 15.8, respectively. Additionally, ChatGPT performed sufficiently in providing educational materials on dermatological diseases, with a 46.94 mean reading ease score. [20]

Conversely, Kianian et al. observed a lower FKGL of ChatGPT's responses (6.3 ± 1.2) than in Bard's responses (10.5 ± 0.8) when asked to generate educational information about uveitis. [21] ChatGPT scored an eighth-grade readability level when generating output responses on radiological cases. [22] Moreover, Xie et al. evaluated the readability of ChatGPT, Bard, and BingAI in generating answers about complex clinical scenarios. Among the three LLMs, ChatGPT had the highest Flesch Reading Ease score. Nonetheless, Bard was a close runner-up, and no significant difference was reported between the two. [18] In summary, although GhatGPT and Google Bard differ significantly in readability, both provide clear, understandable text with a grade level suitable for patients seeking knowledge on PCa.

Almost all the generated answers were stable, except for one question within the "screening and prevention domain." Specifically, when asked, "Should I get screened for prostate cancer?" The 1st answer of ChatGPT was less accurate than the second and third answers. Thus, this question was labeled "inconsistent". It is important to note that only ten questions were tested for stability and compared across the three LLMs, as they are generally stable. In future studies, all inquiries should be tested and objectively evaluated in terms of their accuracy, comprehensiveness, and readability to determine the extent of their stability.

Overall, the steady stream of messages from patients has become a major source of stress in clinics and is one factor that leads to burnout. [23] In the world of medicine, large language models (LLMs), as exemplified by ChatGPT, have demonstrated encouraging possibilities. [24, 25] Furthermore, Haifeng Song et al. demonstrated the extraordinary potential of LLMs in patient education and medical health consultations. [26] Even though they are not yet flawless, LLMs can accurately respond to common queries from PCa patients and can, to a certain extent, analyse certain scenarios. LLMs can be used in patient education and consultation by providing patients with easily understood information on their disease and available treatments, allowing collaborative decision-making. More significantly, LLMs can contribute to the democratization of medical knowledge by providing everyone, regardless of location or socioeconomic background, with fast access to reliable medical information. Particular attention should be given to underprivileged communities living in medical deserts and those having to wait longer for care during pandemics such as the COVID-19 pandemic. Given the speed at which AI is developing, LLMs have limitless potential. [12]

AI chatbots have shown outstanding performance in providing precise, thorough information on PCa. According to Johnson et al., even though there were precedential concerns regarding the ability of ChatGPT to provide information, especially in the context of cancer, their study shed light on the positive capability of ChatGPT in terms of accuracy. Nonetheless, even if AI can learn everything about PCa, it remains an objective source of knowledge since it has never experienced the physical presence of treating such cases. This is described as the knowledge argument theory, in which the physical description of a disease cannot replace the actual perceptual experience of treating it. [27] ChatGPT, like every new invention, raises fear among physicians related to the possibility of replacement. [28] However, there is a fundamental difference between knowing everything about PCa and actually having the experience of treating patients and communicating their needs. Qualia is the philosophical term describing this subjective and personal knowledge gained from physician-patient interactions, the empathy evoked from witnessing patients' suffering, and the tactile feedback experienced during physical examination or surgery. [27] Since these qualia are inaccessible to AI, it is impossible for AI to replace physicians in healthcare education; AI will rather be a valuable assistant if trained adequately. [28]

Limitations

While the study provided promising and insightful results, it had several limitations. First, although incorporating more questions would have clarified statistical differences between the LLMs, this study covered the most relevant, widely asked questions on PCa. Furthermore, ChatGPT retrieves the relevant data from its knowledge base, which is only updated until September 2021. Finally, Google Bard demonstrated a lack of information by refusing to answer one question, which might not have affected the results. These limitations did not affect the reliability of the findings. To our knowledge, this is the first study to compare the performance of ChatGPT and Google Bard in the context of PCa.

Conclusion

In conclusion, ChatGPT and Google Bard performed well in providing informational content on PCa and might be helpful resources for patients and the general public. These study findings emphasize the promising role of AI assistance in improving patients' quality of life and enhancing their education. Future studies should incorporate personalized inquiries and evaluate whether providing additional context would affect the tested outcomes.

Abbreviations

ADT	Androgen Deprivation Therapy
AI	Artificial Intelligence
ASCO	American Society of Clinical Oncology
AUA	American Urological Association
CDC	Centers for Disease Control and Prevention
ChatGPT	Generative Pretrained Transformers
DRE	Digital Rectal Examination
EAU	European Association of Urology
FKGL	Flesch–Kincaid Grade Level
FRE	Flesch Reading Ease
IQR	Interquartile Range
LaMDA	Language Model for Dialogue Applications
LLMs	Large Language Models
NCCN	National Comprehensive Cancer Network
PCa	Prostate Cancer
PCF	Prostate Cancer Foundation
PSA	Prostate-Specific Antigen

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12894-024-01570-0.

Supplementary Material 1

Acknowledgements

None.

Author contributions

A.A. conceptualized the study design, evaluated the answers from LLMs, and revised and edited the manuscript. S.A., N.A., N.A., and F.A. collected the questions, extracted the answers from LLMs, and wrote the original draft of the manuscript. M.A. performed the formal analysis, created the tables and figures, and wrote the original draft of the manuscript. M.A. and B.A. evaluated the answers from LLMs and revised and edited the manuscript.

Funding

None.

Data availability

The datasets used and/or analysed during the current study available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate

None.

Consent for publication

None.

Competing interests

The authors declare no competing interests.

Received: 23 January 2024 / Accepted: 16 August 2024
Published online: 23 August 2024

References

1. Gilson A, et al. How does Chatgpt perform on the United States medical licensing examination? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ. 2023;9:e45312.
2. Miao J, Thongprayoon C, Cheungpasitporn W. Assessing the accuracy of chatgpt on core questions in glomerular disease. Kideny Int Rep. 2023;8:1657–9.
3. Biswas S. Role of chat gpt in public health. Ann Biomed Eng. 2023;51:868–9.
4. Sarraju A, Bruemmer D, Van Iterson E, Cho L, Rodriguez F, Laffin L. Appropriateness of cardiovascular disease prevention recommendations obtained from a popular online chat-based artificial intelligence model. JAMA. 2023;329:842–4.
5. Rawla P. Epidemiology of prostate cancer. World J Oncol. 2019;10:63.
6. Alqahtani WS, et al. Epidemiology of cancer in Saudi Arabia thru 2010–2019: a systematic review with constrained meta-analysis. AIMS Public Health. 2020;7:679.
7. Sekhoacha M, et al. Prostate cancer review: Genetics, diagnosis, treatment options, and alternative approaches. Molecules. 2022;27:5730.
8. Jindal P, MacDermid JC. Assessing reading levels of health information: uses and limitations of flesch formula. Educ Health. 2017;30:84–8.
9. NCCN Guidelines. [cited 2023 Sept 26]. https://www.nccn.org/guidelines/guidelines-detail?category=1&id=1459
10. American Urological Association [Internet]. [cited 2023 Sept 26]. https://www.auanet.org/guidelines-and-quality/guidelines
11. European Association of Urology [Internet]. [cited 2023 Sept 26]. https://uroweb.org/guidelines

12. 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;21:1–4.
13. Pan A et al. Assessment of artificial intelligence chatbot responses to top searched queries about cancer. JAMA Oncol. (2023).
14. Hirosawa T, Mizuta K, Harada Y, Shimizu T. Comparative Evaluation of Diagnostic Accuracy between Google Bard and Physicians. Am J Med. 2023;136:1119–e112318.
15. Lombardo R, Gallo G, Stira J, et al. Quality of information and appropriateness of Open AI outputs for prostate cancer. Prostate Cancer Prostatic Dis. 2024. https://doi.org/10.1038/s41391-024-00789-0.
16. Gebrael G, Sahu KK, Chigarira B, TripathiN, Mathew Thomas V, Sayegh N, Maughan BL, Agarwal N, Swami U, Li H. Enhancing triage efficiency and accuracy in emergency rooms for patients with metastatic prostate cancer: a retrospective analysis of artificial intelligence-assisted triage using ChatGPT 4.0. Cancers. 2023;15(14):3717.
17. Lim ZW et al. Benchmarking large language models' performances for myopia care: a comparative analysis of ChatGPT-3.5, ChatGPT-4.0, and Google Bard. *EBioMedicine*. 95 (2023).
18. Xie Y, Seth I, Hunter-Smith DJ, Rozen WM, Seifman MA. Investigating the impact of innovative AI chatbot on post-pandemic medical education and clinical assistance: a comprehensive analysis. ANZ J Surg. 2023. https://doi.org/10.1111/ans.18666.
19. Cocci A, et al. Quality of information and appropriateness of ChatGPT outputs for urology patients. Prostate Cancer Prostatic Dis. 2023. https://doi.org/10.1038/s41391-023-00705-y.
20. Mondal H, Mondal S, Podder I. Using chatgpt for writing articles for patients' education for dermatological diseases: a pilot study. Indian Dermatol Online J. 2023;14:482–6.
21. Kianian R, Sun D, Crowell EL, Tsui E. The use of large language models to generate education materials about uveitis. Ophthalmol Retina. 2023;23:2468–6530.
22. Kuckelman IJ. Assessing AI-powered patient education: a case study in radiology. Acad Radiol. 2023;23:1076–6332.
23. Liu S, McCoy AB, Wright AP, Carew B, Genkins JZ, Huang SS et al. Leveraging large language models for generating responses to patient messages [Internet]. Cold Spring Harbor Laboratory Press; 2023 [cited 2024 Jan 8]. https://www.medrxiv.org/content/https://doi.org/10.1101/2023.07.14.23292669v1.full-text
24. Johnson SB, King AJ, Warner EL, Aneja S, Kann BH, Bylund CL. Using ChatGPT to evaluate cancer myths and misconceptions: artificial intelligence and cancer information. JNCI cancer Spectr. 2023;7(2):pkad015.
25. Hopkins AM, Logan JM, Kichenadasse G, Sorich MJ. Artificial Intelligence Chatbots will revolutionize how cancer patients access information: CHATGPT represents a paradigm-shift [Internet]. Oxford University Press; 2023 [cited 2024 Jan 8]. https://academic.oup.com/jncics/article/7/2/pkad010/7049531
26. Song H, Xia Y, Luo Z, Liu H, Song Y, Zeng X et al. Evaluating the performance of different large language models on health consultation and patient education in Urolithiasis - Journal of Medical Systems [Internet]. Springer US; 2023 [cited 2024 Jan 8]. https://link.springer.com/article/https://doi.org/10.1007/s10916-023-02021-3
27. Nida-Rümelin M, O Conaill D, Qualia. The knowledge argument [Internet]. Stanford University; 2019 [cited 2023 Oct 24]. https://plato.stanford.edu/entries/qualia-knowledge/#BasiIdea
28. Lombardo R, Cicione A, Santoro G, De Nunzio C. ChatGPT in prostate cancer: myth or reality? Prostate Cancer Prostatic Dis 2023 Nov 10:1–2.

Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.