SIUJ Société Internationale d'Urologie Journal

Article

ChatGPT vs. Gemini: Which Provides Better Information on Bladder Cancer?

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

1 Division of Urology, Department of Surgery, Ministry of National Guard—Health Affairs, Riyadh 11461, Saudi Arabia
2 King Abdullah International Medical Research Center (KAIMRC), Riyadh 11461, Saudi Arabia
3 College of Medicine, King Saud bin Abdulaziz University for Health Sciences, Riyadh 11461, Saudi Arabia
4 Department of Surgical Specialties, College of Medicine, Majmaah University, Majmaah 11461, Saudi Arabia
* Correspondence: seham1alslamh@gmail.com

Abstract: Background/Objectives: Bladder cancer, the most common and heterogeneous malignancy of the urinary tract, presents with diverse types and treatment options, making comprehensive patient education essential. As large language models (LLMs) emerge as a promising resource for disseminating medical information, their accuracy and validity compared to traditional methods remain under-explored. This study aims to evaluate the effectiveness of LLMs in educating the public about bladder cancer. **Methods**: Frequently asked questions regarding bladder cancer were sourced from reputable educational materials and assessed for accuracy, comprehensiveness, readability, and consistency by two independent board-certified urologists, with a third resolving any discrepancies. The study utilized a 3-point Likert scale for accuracy, a 5-point Likert scale for comprehensiveness, and the Flesch–Kincaid (FK) Grade Level and Flesch Reading Ease (FRE) scores to gauge readability. **Results**: ChatGPT-3.5, ChatGPT-4, and Gemini were evaluated on 12 general questions, 6 questions related to diagnosis, 28 concerning treatment, and 7 focused on prevention. Across all categories, the correct response rate was notably high, with ChatGPT-3.5 and ChatGPT-4 achieving 92.5%, compared to 86.3% for Gemini, with no significant difference in accuracy. However, there was a significant difference in comprehensiveness ($p = 0.011$) across the models. Overall, a significant difference in performance was observed among the LLMs ($p < 0.001$), with ChatGPT-4 providing the most college-level responses, though these were the most challenging to read. **Conclusions**: In conclusion, our study adds value to the applications of Artificial Intelligence (AI) in bladder cancer education, with notable insights into the accuracy, comprehensiveness, and stability of the three LLMs.

Keywords: bladder cancer; artificial intelligence; large language models; chatbot; ChatGPT; Gemini

Academic Editor: Peter C. Black

Received: 15 December 2024
Revised: 1 March 2025
Accepted: 3 March 2025
Published: 21 April 2025

Citation: Alasker, A.; Alshathri, N.; Alsalamah, S.; Almansour, N.; Alsalamah, F.; Alghafees, M.; AlKhamees, M.; Alsaikhan, B. ChatGPT vs. Gemini: Which Provides Better Information on Bladder Cancer? *Soc. Int. Urol. J.* **2025**, *6*, 34. https://doi.org/10.3390/siuj6020034

1. Introduction

Bladder cancer is the most common heterogeneous malignancy in the urinary tract, which has variable natural history, types, and treatment options [1]. In Saudi Arabia, the incidence of bladder cancer overall was 1.4 per 100,000 individuals [2]. Additionally, its incidence in Saudi Arabia has increased tenfold in the last decade [3]. The risk factors behind the development of bladder cancer are smoking, exposure to certain chemicals, advancing age, and male gender. Patients usually present with hematuria, whether gross

or microscopic, but require further investigations, such as cystoscopy and urography [4]. The decision on the most beneficial treatment option is based on the accurate staging and grading of the tumor, as well as the type. Non-muscle invasive urothelial bladder cancer is preferably treated with Transurethral Resection of the Bladder Tumor (TURPT), while in muscle-invasive bladder cancer, radical cystectomy, i.e., removal of the bladder, is implemented. Favorable oncological outcomes of bladder cancer treatment, however, require a multidisciplinary approach [5]. Moreover, since it is a disease that results in severe consequences that affect several aspects of life, early detection through screening, in addition to engaging the patient in cancer care through education, are all essential to improve patient outcomes and quality of life [1,6]. For this reason, it has become essential to empower patients to take a proactive approach when it comes to their care by improving the educational resources and tools and raising their awareness about their condition [6].

Artificial Intelligence (AI) systems have gained popularity in recent years for providing information and assisting users online [7]. Moreover, the utilization of the internet for health-related reasons, such as information about cancer, has significantly increased [8,9]. Some bladder cancer patients may utilize the internet as a means to get more information about their condition, and the current technological advancements in large language model (LLM) chatbots have provided another tool from which health information can be obtained [10,11]. Chat Generative Pretrained Transformer (ChatGPT), which was recently developed by the company OpenAI, has captured the attention of many with its ability to provide information and build conversations with users due to its deep learning programming. ChatGPT provides answers in a contextualized manner that are relevant to the input of the user. These features made the application of ChatGPT as a source of medical information a reality [12]. Another emerging AI chatbot that was released by Google is Bard, which is powered by Language Model for Dialogue Applications (LaMDA) and incorporates language recognition and conversation processing while interfacing with Google search [13].

This new era has raised both hopes for the future applications of these tools in healthcare as well as concerns about their current quality and accuracy of responses that deliver critical medical information, especially when it comes to cancer [9,14]. Furthermore, monitoring these platforms' outputs for accuracy and their ability to keep up with the changing nature of medical sciences and its new and increasing advancements is a necessity, given the importance of this information in the cancer research field and to healthcare communicators [9].

However, limited studies have established data about the quality and accuracy of information that these chatbots provide to cancer patients. Therefore, the aim of this study is to assess the effectiveness of large language models (LLMs), ChatGPT (third and fourth generations), and Gemini in providing bladder cancer patients with information about their condition and evaluate the quality of the answers provided to these patients.

2. Materials and Methods

Frequently asked questions from various trusted educational websites were compiled into an Excel file. The criteria on which these questions were selected are (1) questions frequently asked by patients and the public and (2) questions that target general disease knowledge, treatment, diagnosis, and prevention. The distribution of questions reflects the natural emphasis found in widely available educational resources, ensuring that the most relevant and commonly searched topics are well represented. These questions were then reviewed by board-certified urologists for optimum selection. A complete list of the questions, along with their respective sources, is available in the supplementary material for reference. The questions were then inputted into two LLMs

(ChatGPT-3.5, ChatGPT-4, and Gemini), all available at https://chat.openai.com/chat and https://gemini.google.com/app websites (accessed on 27 March 2024). The quality of responses was assessed based on accuracy, comprehensiveness, patient readability, and stability. A 3-point scale was employed to measure accuracy, with one indicating correct information, two indicating a mix of correct and incorrect, and three indicating completely incorrect information. For assessing the comprehensiveness of responses, a 5-point Likert scale was used, where one represented "very comprehensive" and five represented "very inadequate". As for readability, the output answers, sentences, words, syllables per word, and words per sentence were the factors analyzed. 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/ (accessed on 27 March 2024). A higher Flesch Reading Ease Score signifies that the text is easy to read, whereas the Flesch–Kincaid Grade Level indicates the educational grade level required to comprehend the text. The stability of the output text was analyzed due to the various responses generated for the same question by the LLMs. Stability was assessed by two independent reviewers who subjectively evaluated whether the second and third answers were accurate compared to the first generated answer. Three responses were generated for each question, with the chat history cleared after each trial. The first answers to these questions were evaluated separately by two board-certified urologists, all referring to the same resource for accuracy and comprehension. All responses were evaluated using guidelines from the American Urological Association (AUA), Canadian Urological Association (CUA), National Comprehensive Cancer Network (NCCN), and European Association of Urology (EAU) to ensure consistency [15–18]. Any discrepancies in the evaluation were then independently resolved by a blinded third board-certified urologist.

Statistical Analysis

Statistical analysis was conducted using RStudio software (version 4.3.1). Descriptive statistics were employed to summarize the characteristics of different large language models (LLMs) and their performance across various categories related to bladder cancer. Categorical variables were expressed as frequencies and percentages. To assess the association between LLMs and categorical variables, Fisher's exact test was utilized. Additionally, continuous variables, such as grade-level scores, were presented as medians with interquartile ranges (IQRs), and the Kruskal–Wallis rank sum test was applied to evaluate significant differences among LLMs. All statistical tests were two-tailed, and p-values less than 0.05 were considered statistically significant.

3. Results

In the current study, three large language models (LLMs), including ChatGPT-3.5, ChatGPT-4, and Gemini, were assessed using 12 general questions, 6 diagnosis-related questions, 28 treatment-related questions, and 7 prevention-related questions, which accounted for 22.6%, 11.3%, 52.8%, and 13.2%, respectively (Figure 1).

Figure 2 presents the analysis of the accuracy of different LLMs in providing information about bladder cancer. The overall analysis of 53 questions revealed no significant differences among the LLMs ($p = 0.655$). Across all categories, the proportion of correct responses was high, with 92.5% for ChatGPT-3.5 and ChatGPT-4 and 86.3% for Gemini. There were no significant differences in the proportions of correct answers among the three LLMs in terms of the questions of general domains ($p > 0.999$), diagnosis ($p > 0.999$), treatment ($p = 0.848$), and prevention ($p = 0.079$).

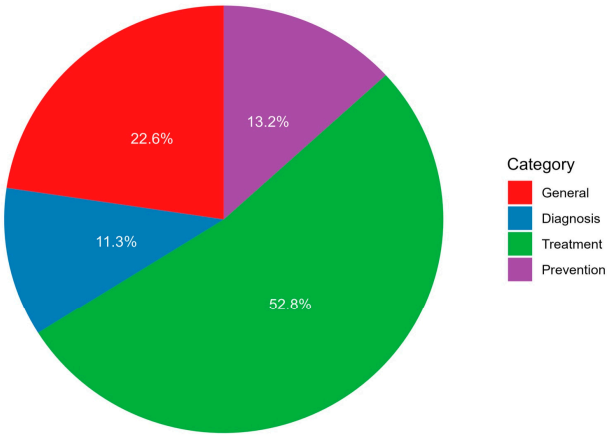

Figure 1. Distribution of question categories in bladder cancer education.

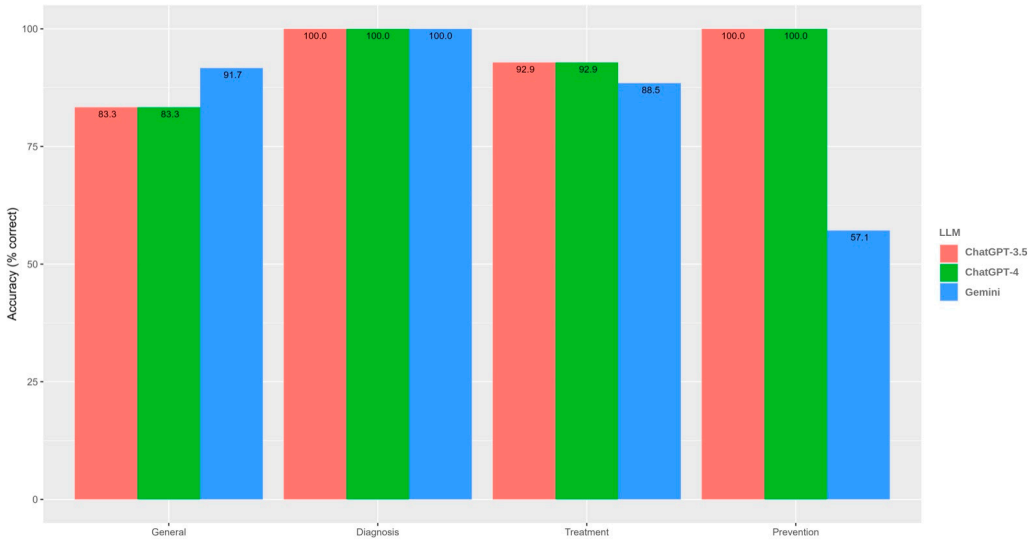

Figure 2. Accuracy of responses by large language models (LLMs) across different question categories.

The overall analysis revealed a significant difference among the LLMs in terms of their comprehensiveness ($p = 0.011$). Comprehensive and very comprehensive responses were provided for 75.4% of questions by ChatGPT-3.5, 83.0% by ChatGPT-4, and 68.6% by Gemini. Additionally, in the treatment category, ChatGPT-3.5 displayed a higher combined proportion (75.0%) compared to ChatGPT-4 (78.6%) and Gemini (57.7%), and the difference was statistically significant ($p = 0.007$, Table 1). No other question categories showed statistically significant differences among the LLMs.

Table 2 presents an analysis of grade-level scores among different large language models (LLMs) regarding bladder cancer information. The overall comparison revealed a significant difference among the LLMs ($p < 0.001$), with college-level responses apparent in 34.0% of the questions by Gemini, 69.8% by ChatGPT-3.5, and 75.5% by ChatGPT-4. Furthermore, in the treatment category, ChatGPT-3.5 had a proportion of 71.4%, ChatGPT-4 had 75.0%, and Gemini had 39.3% categorized as college-level questions. This discrepancy was statistically significant ($p < 0.001$). In general questions, Gemini had significantly lower proportions of college-level questions (16.7%) compared to ChatGPT-3.5 (66.7%) and ChatGPT-4 (83.3%). In the diagnosis category, ChatGPT-3.5 had a proportion of 66.7%, ChatGPT-4 had 83.3%, and Gemini had 16.7% categorized as "College", although this difference was borderline significant ($p = 0.050$, Table 2).

Table 1. Analysis of the comprehensiveness of different large language models (LLMs).

Characteristic	Missing	ChatGPT	ChatGPT Plus	Gemini	*p*-Value
Overall (*n* = 53)	2 (1.3%)				0.011
Very inadequate		5 (9.4%)	3 (5.7%)	5 (9.8%)	
Inadequate		1 (1.9%)	1 (1.9%)	4 (7.8%)	
Neither comprehensive nor inadequate		7 (13.2%)	5 (9.4%)	7 (13.7%)	
Comprehensive		35 (66.0%)	22 (41.5%)	22 (43.1%)	
Very comprehensive		5 (9.4%)	22 (41.5%)	13 (25.5%)	
General (*n* = 12)	0 (0%)				0.291
Very inadequate		0 (0.0%)	0 (0.0%)	0 (0.0%)	
Inadequate		1 (8.3%)	0 (0.0%)	0 (0.0%)	
Neither comprehensive nor inadequate		3 (25.0%)	0 (0.0%)	2 (16.7%)	
Comprehensive		7 (58.3%)	8 (66.7%)	6 (50.0%)	
Very comprehensive		1 (8.3%)	4 (33.3%)	4 (33.3%)	
Diagnosis (*n* = 6)	0 (0%)				>0.999
Very inadequate		0 (0.0%)	0 (0.0%)	0 (0.0%)	
Inadequate		0 (0.0%)	0 (0.0%)	1 (16.7%)	
Neither comprehensive nor inadequate		1 (16.7%)	1 (16.7%)	1 (16.7%)	
Comprehensive		2 (33.3%)	2 (33.3%)	1 (16.7%)	
Very comprehensive		3 (50.0%)	3 (50.0%)	3 (50.0%)	
Treatment (*n* = 28)	2 (2.4%)				0.007
Very inadequate		5 (17.9%)	3 (10.7%)	5 (19.2%)	
Inadequate		0 (0.0%)	1 (3.6%)	3 (11.5%)	
Neither comprehensive nor inadequate		2 (7.1%)	2 (7.1%)	3 (11.5%)	
Comprehensive		20 (71.4%)	10 (35.7%)	10 (38.5%)	
Very comprehensive		1 (3.6%)	12 (42.9%)	5 (19.2%)	
Prevention (*n* = 7)	0 (0%)				0.205
Very inadequate		0 (0.0%)	0 (0.0%)	0 (0.0%)	
Inadequate		0 (0.0%)	0 (0.0%)	0 (0.0%)	
Neither comprehensive nor inadequate		1 (14.3%)	2 (28.6%)	1 (14.3%)	
Comprehensive		6 (85.7%)	2 (28.6%)	5 (71.4%)	
Very comprehensive		0 (0.0%)	3 (42.9%)	1 (14.3%)	

n (%). Fisher's exact test.

Table 2. Analysis of the grade-level score of different large language models (LLMs).

Characteristic	Missing	ChatGPT	ChatGPT Plus	Gemini	*p*-Value
Overall (*n* = 53)	0 (0%)				<0.001
6th grade		0 (0.0%)	0 (0.0%)	1 (1.9%)	
7th grade		0 (0.0%)	0 (0.0%)	4 (7.5%)	
8th & 9th grade		2 (3.8%)	1 (1.9%)	13 (24.5%)	
10th to 12th grade		8 (15.1%)	4 (7.5%)	17 (32.1%)	
College		37 (69.8%)	40 (75.5%)	18 (34.0%)	
College graduate		6 (11.3%)	7 (13.2%)	0 (0.0%)	
Professional		0 (0.0%)	1 (1.9%)	0 (0.0%)	
General (*n* = 12)	0 (0%)				0.016
6th grade		0 (0.0%)	0 (0.0%)	0 (0.0%)	
7th grade		0 (0.0%)	0 (0.0%)	2 (16.7%)	
8th & 9th grade		1 (8.3%)	1 (8.3%)	3 (25.0%)	
10th to 12th grade		3 (25.0%)	1 (8.3%)	5 (41.7%)	

Table 2. *Cont.*

Characteristic	Missing	ChatGPT	ChatGPT Plus	Gemini	*p*-Value
College		8 (66.7%)	10 (83.3%)	2 (16.7%)	
College graduate		0 (0.0%)	0 (0.0%)	0 (0.0%)	
Professional		0 (0.0%)	0 (0.0%)	0 (0.0%)	
Diagnosis (*n* = 6)	0 (0%)				0.050
6th grade		0 (0.0%)	0 (0.0%)	0 (0.0%)	
7th grade		0 (0.0%)	0 (0.0%)	0 (0.0%)	
8th & 9th grade		0 (0.0%)	0 (0.0%)	2 (33.3%)	
10th to 12th grade		1 (16.7%)	0 (0.0%)	3 (50.0%)	
College		4 (66.7%)	5 (83.3%)	1 (16.7%)	
College graduate		1 (16.7%)	1 (16.7%)	0 (0.0%)	
Professional		0 (0.0%)	0 (0.0%)	0 (0.0%)	
Treatment (*n* = 28)	0 (0%)				<0.001
6th grade		0 (0.0%)	0 (0.0%)	1 (3.6%)	
7th grade		0 (0.0%)	0 (0.0%)	2 (7.1%)	
8th & 9th grade		1 (3.6%)	0 (0.0%)	7 (25.0%)	
10th to 12th grade		2 (7.1%)	2 (7.1%)	7 (25.0%)	
College		20 (71.4%)	21 (75.0%)	11 (39.3%)	
College graduate		5 (17.9%)	4 (14.3%)	0 (0.0%)	
Professional		0 (0.0%)	1 (3.6%)	0 (0.0%)	
Prevention (*n* = 7)	0 (0%)				0.561
6th grade		0 (0.0%)	0 (0.0%)	0 (0.0%)	
7th grade		0 (0.0%)	0 (0.0%)	0 (0.0%)	
8th & 9th grade		0 (0.0%)	0 (0.0%)	1 (14.3%)	
10th to 12th grade		2 (28.6%)	1 (14.3%)	2 (28.6%)	
College		5 (71.4%)	4 (57.1%)	4 (57.1%)	
College graduate		0 (0.0%)	2 (28.6%)	0 (0.0%)	
Professional		0 (0.0%)	0 (0.0%)	0 (0.0%)	

n (%). Fisher's exact test.

Table 3 illustrates the analysis of the reading notes of different large language models (LLMs). The overall comparison revealed a significant difference among the LLMs ($p < 0.001$), with difficulty in reading for 69.8% of responses provided by ChatGPT-3.5, 75.5% of ChatGPT-4, and 34.0% of Gemini. Notably, in the treatment category, ChatGPT-3.5 had 71.4%, ChatGPT-4 had 75.0%, and Gemini had 39.3% categorized as "Difficult to read", indicating a substantial discrepancy among them ($p < 0.001$). Similarly, in the diagnosis category, ChatGPT-3.5 had 66.7%, ChatGPT-4 had 83.3%, and Gemini had 16.7% categorized as "Difficult to read", and the difference was statistically significant ($p = 0.019$, Table 3).

Table 3. Analysis of the reading notes of different large language models (LLMs).

Characteristic	Missing	ChatGPT	ChatGPT Plus	Gemini	*p*-Value
Overall (*n* = 53)	0 (0%)				<0.001
Plain English		2 (3.8%)	1 (1.9%)	13 (24.5%)	
Fairly easy to read		0 (0.0%)	0 (0.0%)	4 (7.5%)	
Easy to read		0 (0.0%)	0 (0.0%)	1 (1.9%)	
Difficult to read		37 (69.8%)	40 (75.5%)	18 (34.0%)	
Fairly difficult to read		8 (15.1%)	4 (7.5%)	17 (32.1%)	
Very difficult to read		6 (11.3%)	7 (13.2%)	0 (0.0%)	
Extremely difficult to read		0 (0.0%)	1 (1.9%)	0 (0.0%)	

Table 3. *Cont.*

Characteristic	Missing	ChatGPT	ChatGPT Plus	Gemini	*p*-Value
General (*n* = 12)	0 (0%)				0.019
Plain English		1 (8.3%)	1 (8.3%)	3 (25.0%)	
Fairly easy to read		0 (0.0%)	0 (0.0%)	2 (16.7%)	
Easy to read		0 (0.0%)	0 (0.0%)	0 (0.0%)	
Difficult to read		8 (66.7%)	10 (83.3%)	2 (16.7%)	
Fairly difficult to read		3 (25.0%)	1 (8.3%)	5 (41.7%)	
Very difficult to read		0 (0.0%)	0 (0.0%)	0 (0.0%)	
Extremely difficult to read		0 (0.0%)	0 (0.0%)	0 (0.0%)	
Diagnosis (*n* = 6)	0 (0%)				0.055
Plain English		0 (0.0%)	0 (0.0%)	2 (33.3%)	
Fairly easy to read		0 (0.0%)	0 (0.0%)	0 (0.0%)	
Easy to read		0 (0.0%)	0 (0.0%)	0 (0.0%)	
Difficult to read		4 (66.7%)	5 (83.3%)	1 (16.7%)	
Fairly difficult to read		1 (16.7%)	0 (0.0%)	3 (50.0%)	
Very difficult to read		1 (16.7%)	1 (16.7%)	0 (0.0%)	
Extremely difficult to read		0 (0.0%)	0 (0.0%)	0 (0.0%)	
Treatment (*n* = 28)	0 (0%)				<0.001
Plain English		1 (3.6%)	0 (0.0%)	7 (25.0%)	
Fairly easy to read		0 (0.0%)	0 (0.0%)	2 (7.1%)	
Easy to read		0 (0.0%)	0 (0.0%)	1 (3.6%)	
Difficult to read		20 (71.4%)	21 (75.0%)	11 (39.3%)	
Fairly difficult to read		2 (7.1%)	2 (7.1%)	7 (25.0%)	
Very difficult to read		5 (17.9%)	4 (14.3%)	0 (0.0%)	
Extremely difficult to read		0 (0.0%)	1 (3.6%)	0 (0.0%)	
Prevention (*n* = 7)	0 (0%)				0.562
Plain English		0 (0.0%)	0 (0.0%)	1 (14.3%)	
Fairly easy to read		0 (0.0%)	0 (0.0%)	0 (0.0%)	
Easy to read		0 (0.0%)	0 (0.0%)	0 (0.0%)	
Difficult to read		5 (71.4%)	4 (57.1%)	4 (57.1%)	
Fairly difficult to read		2 (28.6%)	1 (14.3%)	2 (28.6%)	
Very difficult to read		0 (0.0%)	2 (28.6%)	0 (0.0%)	
Extremely difficult to read		0 (0.0%)	0 (0.0%)	0 (0.0%)	

n (%). Fisher's exact test.

There were notable variations observed among LLMs for the number of words, with ChatGPT-3.5 showing a median of 207.0 (IQR = 166.0 to 240.0), ChatGPT-4 with 295.0 (IQR = 232.0 to 342.0), and Gemini with 274.0 (IQR = 223.0 to 341.0), indicating an increasing trend from ChatGPT-3.5 to Gemini to ChatGPT-4 ($p < 0.001$). Similarly, an increasing trend was observed for sentences, syllables, and syllable/word ratio, with ChatGPT-4 consistently showing the highest values, followed by Gemini and then ChatGPT-3.5. For instance, the median number of sentences was 10.0 (IQR = 7.0 to 13.0) for ChatGPT-3.5, 18.0 (IQR = 10.0 to 21.0) for ChatGPT-4, and 15.0 (IQR = 11.0 to 21.0) for Gemini ($p < 0.001$). Regarding the Flesch Reading Ease (FRE) Score, ChatGPT-3.5 had a median of 43.4 (IQR = 34.2 to 48.0), ChatGPT-4 had 40.3 (IQR = 35.0 to 44.9), and Gemini had 54.3 (IQR = 47.4 to 61.7). However, no significant difference was observed in the Flesch–Kincaid (FK) Reading Level among the LLMs ($p = 0.093$, Table 4 and Figure 3).

Table 4. A description of selected numerical parameters of large language models (LLMs), including words, sentences, syllables, word/sentence, syllable/word, Flesch Reading Ease (FRE) score, and FK Reading Levels.

Characteristic	Missing	ChatGPT	ChatGPT Plus	Gemini	*p*-Value
Words	0 (0%)	207.0 (166.0–240.0)	295.0 (232.0–342.0)	274.0 (223.0–341.0)	<0.001
Sentences	0 (0%)	10.0 (7.0–13.0)	18.0 (10.0–21.0)	15.0 (11.0–21.0)	<0.001
Syllables	0 (0%)	337.0 (285.0–404.0)	507.0 (390.0–601.0)	427.0 (351.0–537.0)	<0.001
Word/sentence	0 (0%)	20.0 (17.1–23.5)	18.6 (15.3–22.1)	17.3 (16.1–21.1)	0.099
Syllable/word	0 (0%)	1.7 (1.6–1.8)	1.8 (1.7–1.8)	1.6 (1.5–1.6)	<0.001
FRE Score	0 (0%)	43.4 (34.2–48.0)	40.3 (35.0–44.9)	54.3 (47.4–61.7)	<0.001
FK Reading Level	0 (0%)	13.6 (11.7–15.4)	12.6 (11.5–13.8)	11.0 (9.4–45,116.0)	0.093

IQR: interquartile range. Median (IQR). Kruskal-Wallis rank sum test.

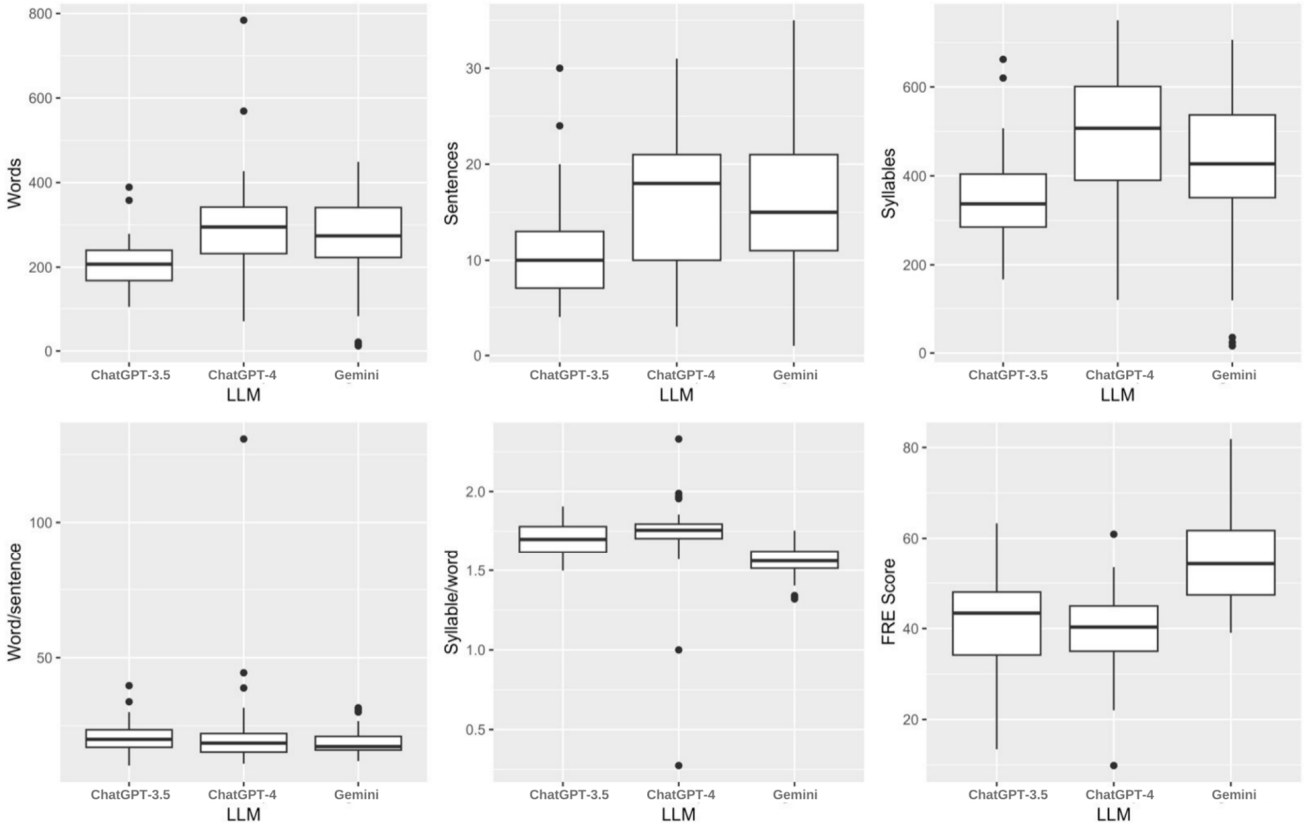

Figure 3. Readability metrics of responses for large language models (LLMs).

The stability analysis was performed on 10 questions, among which three questions were related to diagnosis, another three questions were related to treatment, and four questions were related to prevention. Results showed no significant differences in stability between the three LLMs under investigation for the overall questions and for each subscale (Table 5).

Table 5. Analysis of the stability of different large language models (LLMs).

Characteristic	ChatGPT	ChatGPT Plus	Gemini	*p*-Value
Overall (*n* = 10)				>0.999
Consistent	9 (90.0%)	9 (90.0%)	8 (80.0%)	
Inconsistent	1 (10.0%)	1 (10.0%)	2 (20.0%)	

Table 5. *Cont.*

Characteristic	ChatGPT	ChatGPT Plus	Gemini	*p*-Value
Diagnosis (n = 3)				0.671
Consistent	3 (100.0%)	2 (66.7%)	1 (33.3%)	
Inconsistent	0 (0.0%)	1 (33.3%)	2 (66.7%)	
Treatment (n = 3)				>0.999
Consistent	2 (66.7%)	3 (100.0%)	3 (100.0%)	
Inconsistent	1 (33.3%)	0 (0.0%)	0 (0.0%)	
Prevention (n = 4)				
Consistent	4 (100.0%)	4 (100.0%)	4 (100.0%)	NA
Inconsistent	0 (0.0%)	0 (0.0%)	0 (0.0%)	

NA: non-applicable.

4. Discussion

Our study focused on evaluating the performance of two LLMs in answering questions related to bladder cancer. While previous studies have reported that ChatGPT provides unsatisfactory results regarding bladder cancer information, our findings demonstrated promising accuracy rates, consistent stability, and varying levels of comprehensiveness among the three LLMs [19,20].

LLMs have demonstrated high accuracy in providing information about bladder cancer, underscoring their potential as valuable resources in the medical field. Our findings align with those of Ozgor et al., who reported that ChatGPT's responses to questions about urological cancers exhibited high accuracy rates. However, when evaluated against the EAU 2023 Guidelines, Ozgor et al. found ChatGPT's performance to be inadequate [19]. Additionally, ChatGPT's responses to various urologic conditions, including bladder cancer, were generally well balanced, though treatment-related answers only achieved a moderate quality score on the DISCERN questionnaire, scoring 3 out of 5 points [20]. Consistent with these findings, our study also identified that most inaccuracies were related to bladder cancer treatment. Similarly, Musheyev et al. assessed the quality and accuracy of information about urological cancers provided by four AI chatbots (ChatGPT, Perplexity, ChatSonic, and Microsoft Bing AI) using the top five Google Trends search queries. While these AI chatbots generally delivered accurate and moderately high-quality information, they often lacked clear, actionable instructions [21].

Although all LLMs delivered highly accurate responses regarding bladder cancer information, instances of errors were observed, which stemmed from factors such as misinterpretation of ambiguous questions, reliance on outdated or incorrect sources, and a lack of contextual understanding of nuanced medical topics. For example, when asked, "How will we know if the treatment of bladder cancer is working?", all three LLMs incorrectly recommended blood and urine tests, CT and MRI, cystoscopy, and biopsy, which do not fully align with the latest guidelines. These errors, despite the overall high accuracy rates, underscore the importance of critical evaluation by medical professionals. While LLMs show great promise in providing accurate information about bladder cancer, it is essential to recognize and address their limitations and potential errors to ensure their safe and effective use in a medical context.

It is evident how crucial it is for readers to be able to accurately read, comprehend, and apply information about their condition. Literature has reported that understanding AI-based LLM results can be challenging, with some findings indicating that reading and comprehending this content requires adequate training [22–24]. Our study demonstrated that Gemini provided more easily readable answers compared to both ChatGPT-3.5 and ChatGPT-4. Multiple studies have confirmed that ChatGPT frequently produces responses at a post-secondary, particularly college, grade level [22,25,26]. The higher

percentage of "difficult to read" responses in ChatGPT can be attributed to its tendency to generate complex sentence structures, the use of advanced vocabulary, and the inclusion of detailed contextual explanations that may surpass the reading level of the general public. Additionally, as shown by Abou-Abdallah et al., ChatGPT's readability was considered poor, with a mean FRES value of 38.9, placing it in the fairly difficult category [27]. A study assessing the readability of AI chatbot responses to the 100 most frequently asked questions about cardiopulmonary resuscitation found that Gemini's responses were the easiest to read, while ChatGPT's were the most challenging [28]. Regarding the stability and reproducibility of the responses, all the answers generated were consistent. However, since only ten questions were evaluated for stability and compared across the three LLMs, it is important to test all research questions in future studies to accurately determine their stability.

The use of LLMs in patient education raises important ethical concerns, particularly regarding trust, responsibility, and the potential for misinformation. While these models can enhance accessibility to health information, patients may over-rely on AI-generated responses without consulting healthcare professionals. Ensuring transparency, accuracy, and alignment with evidence-based guidelines is essential to mitigate risks. Future efforts should focus on integrating expert oversight, improving source attribution, and developing safeguards to prevent the dissemination of misleading or outdated information.

Limitations

This study has offered valuable insights into the potential integration of LLM chatbots in health education. However, certain limitations should be acknowledged and addressed in future research. For instance, ChatGPT's knowledge base is only updated until September 2021, which may limit the relevance of its responses. Additionally, although independent evaluation was conducted to ensure blinding to the type of LLM, factors such as the formatting of the answers could inadvertently reveal the identity of specific LLMs. One exciting avenue for future research is the inclusion of real-world user testing. While this study evaluated LLM responses in a controlled setting, assessing how actual patients interact with and comprehend this information would provide even deeper insights into their practical usability and impact. Moreover, while accuracy was the primary focus, a key opportunity lies in assessing whether patients can translate this information into actionable health decisions. Investigating the real-world applicability of LLM-generated responses would provide deeper insights into their role in empowering patient education and decision-making. Despite these limitations, the study remains reliable. Future research should consider including other LLMs and focus on crafting contextualized questions that closely mimic real-world scenarios.

5. Conclusions

In conclusion, the application of AI in bladder cancer education is steadily emerging with promising potential. Our study contributes valuable insights into this field by highlighting the accuracy, comprehensiveness, and stability of the three LLMs in answering bladder cancer-related inquiries. While these AI-driven tools can enhance patient education, they should be used as a supplement rather than a replacement for professional medical advice. Future research is essential to address these challenges, refine LLM performance, and ensure their safe and effective integration into clinical practice.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/siuj6020034/s1, Table S1: List of Questions Used in Each Category and Sources of Information.

Author Contributions: Conceptualization, A.A., N.A. (Nada Alshathri), S.A., N.A. (Nura Almansour), F.A., M.A. (Mohammad Alghafees), M.A. (Mohammad AlKhamees) and B.A.; Data Curation, A.A., N.A. (Nada Alshathri), S.A., N.A. (Nura Almansour), F.A., M.A. (Mohammad AlKhamees) and B.A.; Formal Analysis, A.A., N.A. (Nada Alshathri), S.A., N.A. (Nura Almansour), F.A., M.A. (Mohammad Alghafees), M.A. (Mohammad AlKhamees) and B.A.; Investigation, A.A., N.A. (Nada Alshathri), S.A., N.A. (Nura Almansour), F.A., M.A. (Mohammad Alghafees), M.A. (Mohammad AlKhamees) and B.A.; Methodology, A.A., N.A. (Nada Alshathri), S.A., N.A. (Nura Almansour), F.A., M.A. (Mohammad Alghafees), M.A. (Mohammad AlKhamees) and B.A.; Project Administration, A.A., S.A. and M.A. (Mohammad Alghafees); Resources, A.A., N.A. (Nada Alshathri), S.A., N.A. (Nura Almansour), F.A., M.A. (Mohammad Alghafees), M.A. (Mohammad AlKhamees) and B.A.; Software, A.A., N.A. (Nada Alshathri), S.A., N.A. (Nura Almansour), F.A., M.A. (Mohammad Alghafees), M.A. (Mohammad AlKhamees) and B.A.; Supervision, A.A., N.A. (Nada Alshathri), S.A., N.A. (Nura Almansour), F.A., M.A. (Mohammad Alghafees), M.A. (Mohammad AlKhamees) and B.A.; Validation, A.A., N.A. (Nada Alshathri), S.A., N.A. (Nura Almansour), F.A., M.A. (Mohammad Alghafees), M.A. (Mohammad AlKhamees) and B.A.; Visualization, A.A., N.A. (Nada Alshathri), S.A., N.A. (Nura Almansour), F.A., M.A. (Mohammad Alghafees), M.A. (Mohammad AlKhamees) and B.A.; Writing—Original Draft, N.A. (Nada Alshathri), S.A., N.A. (Nura Almansour) and F.A.; Writing—Review and Editing, A.A., N.A. (Nada Alshathri), S.A., N.A. (Nura Almansour), F.A., M.A. (Mohammad Alghafees), M.A. (Mohammad AlKhamees) and B.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data supporting this study are not publicly available due to ethical and privacy considerations.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Kirkali, Z.; Chan, T.; Manoharan, M.; Algaba, F.; Busch, C.; Cheng, L.; Kiemeney, L.; Kriegmair, M.; Montironi, R.; Murphy, W.M.; et al. Bladder cancer: Epidemiology, staging and grading, and diagnosis. *Urology* **2005**, *66* (Suppl. 1), 4–34. [CrossRef] [PubMed]
2. Alghafees, M.A.; Alqahtani, M.A.; Musalli, Z.F.; Alasker, A. Bladder cancer in Saudi Arabia: A registry-based nationwide descriptive epidemiological and survival analysis. *Ann. Saudi Med.* **2022**, *42*, 17–28. [CrossRef] [PubMed] [PubMed Central]
3. Althubiti, M.A.; Nour Eldein, M.M. Trends in the incidence and mortality of cancer in Saudi Arabia. *Saudi Med. J.* **2018**, *39*, 1259–1262. [CrossRef] [PubMed] [PubMed Central]
4. Lenis, A.T.; Lec, P.M.; Chamie, K.; Mshs, M.D. Bladder Cancer: A Review. *JAMA* **2020**, *324*, 1980–1991. [CrossRef] [PubMed]
5. Stein, J.P.; Lieskovsky, G.; Cote, R.; Groshen, S.; Feng, A.C.; Boyd, S.; Skinner, E.; Bochner, B.; Thangathurai, D.; Mikhail, M.; et al. Radical cystectomy in the treatment of invasive bladder cancer: Long-term results in 1054 patients. *J. Clin. Oncol.* **2001**, *19*, 666–675. [CrossRef]
6. Quale, D.Z.; Bangs, R.; Smith, M.; Guttman, D.; Northam, T.; Winterbottom, A.; Necchi, A.; Fiorini, E.; Demkiw, S. Bladder Cancer Patient Advocacy: A Global Perspective. *Bladder Cancer* **2015**, *1*, 117–122. [CrossRef] [PubMed] [PubMed Central]
7. Miner, A.S.; Laranjo, L.; Kocaballi, A.B. Chatbots in the fight against the COVID-19 pandemic. *NPJ Digit. Med.* **2020**, *3*, 65. [CrossRef] [PubMed]
8. Calixte, R.; Rivera, A.; Oridota, O.; Beauchamp, W.; Camacho-Rivera, M. Social and demographic patterns of health-related Internet use among adults in the United States: A secondary data analysis of the health information national trends survey. *Int. J. Environ. Res. Public Health* **2020**, *17*, 6856. [CrossRef]
9. Johnson, S.B.; King, A.J.; Warner, E.L.; Aneja, S.; Kann, B.H.; Bylund, C.L. Using ChatGPT to evaluate cancer myths and misconceptions: Artificial intelligence and cancer information. *JNCI Cancer Spectr.* **2023**, *7*, pkad015. [CrossRef]
10. Corfield, J.M.; Abouassaly, R.; Lawrentschuk, N. Health information quality on the internet for bladder cancer and urinary diversion: A multi-lingual analysis. *Minerva Urol. E Nefrol.=Ital. J. Urol. Nephrol.* **2017**, *70*, 137–143. [CrossRef]
11. Shahsavar, Y.; Choudhury, A. User Intentions to Use ChatGPT for Self-Diagnosis and Health-Related Purposes: Cross-sectional Survey Study. *JMIR Human Factors* **2023**, *10*, e47564. [CrossRef] [PubMed]

12. Dave, T.; Athaluri, S.A.; Singh, S. ChatGPT in medicine: An overview of its applications, advantages, limitations, future prospects, and ethical considerations. *Front. Artif. Intell.* **2023**, *6*, 1169595. [CrossRef]
13. King, M.R. Can Bard, Google's Experimental Chatbot Based on the LaMDA Large Language Model, Help to Analyze the Gender and Racial Diversity of Authors in Your Cited Scientific References? *Cell. Mol. Bioeng.* **2023**, *16*, 175–179. [CrossRef]
14. Koski, E.; Murphy, J. AI in Healthcare. In *Nurses and Midwives in the Digital Age*; IOS Press: Bristol, UK, 2021.
15. American Urological Association [Internet]. Available online: https://www.auanet.org/guidelines-and-quality/guidelines (accessed on 28 February 2025).
16. Canadian Urological Association [Internet]. Available online: https://www.cua.org/guidelines (accessed on 28 February 2025).
17. NCCN Guidelines. Available online: https://www.nccn.org/guidelines/guidelines-detail?category=1&id=1459 (accessed on 28 February 2025).
18. European Association of Urology [Internet]. Available online: https://uroweb.org/guidelines (accessed on 28 February 2025).
19. Ozgor, F.; Caglar, U.; Halis, A.; Cakir, H.; Aksu, U.C.; Ayranci, A.; Sarilar, O. Urological Cancers and ChatGPT: Assessing the Quality of Information and Possible Risks for Patients. *Clin. Genitourin. Cancer* **2024**, *22*, 454–457.e4. [CrossRef] [PubMed]
20. Szczesniewski, J.J.; Tellez Fouz, C.; Ramos Alba, A.; Diaz Goizueta, F.J.; García Tello, A.; Llanes González, L. ChatGPT and most frequent urological diseases: Analysing the quality of information and potential risks for patients. *World J. Urol.* **2023**, *41*, 3149–3153. [CrossRef] [PubMed]
21. Musheyev, D.; Pan, A.; Loeb, S.; Kabarriti, A.E. How Well Do Artificial Intelligence Chatbots Respond to the Top Search Queries About Urological Malignancies? *Eur. Urol.* **2024**, *85*, 13–16. [CrossRef] [PubMed]
22. Davis, R.; Eppler, M.; Ayo-Ajibola, O.; Loh-Doyle, J.C.; Nabhani, J.; Samplaski, M.; Gill, I.; Cacciamani, G.E. Evaluating the Effectiveness of Artificial Intelligence-powered Large Language Models Application in Disseminating Appropriate and Readable Health Information in Urology. *J. Urol.* **2023**, *210*, 688–694. [CrossRef] [PubMed]
23. Momenaei, B.; Wakabayashi, T.; Shahlaee, A.; Durrani, A.F.; Pandit, S.A.; Wang, K.; Mansour, H.A.; Abishek, R.M.; Xu, D.; Sridhar, J.; et al. Appropriateness and Readability of ChatGPT-4-Generated Responses for Surgical Treatment of Retinal Diseases. *Ophthalmol. Retin.* **2023**, *7*, 862–868. [CrossRef] [PubMed]
24. Robinson, M.A.; Belzberg, M.; Thakker, S.; Bibee, K.; Merkel, E.; MacFarlane, D.F.; Lim, J.; Scott, J.F.; Deng, M.; Lewin, J.; et al. Assessing the accuracy, usefulness, and readability of artificial-intelligence-generated responses to common dermatologic surgery questions for patient education: A double-blinded comparative study of ChatGPT and Google Bard. *J. Am. Acad. Dermatol.* **2024**, *90*, 1078–1080. [CrossRef] [PubMed]
25. Hershenhouse, J.S.; Mokhtar, D.; Eppler, M.B.; Rodler, S.; Storino Ramacciotti, L.; Ganjavi, C.; Hom, B.; Davis, R.J.; Tran, J.; Russo, G.I.; et al. Accuracy, readability, and understandability of large language models for prostate cancer information to the public. *Prostate Cancer Prostatic Dis.* **2024**, 1–6. [CrossRef] [PubMed]
26. Zaleski, A.L.; Berkowsky, R.; Craig, K.J.T.; Pescatello, L.S. Comprehensiveness, Accuracy, and Readability of Exercise Recommendations Provided by an AI-Based Chatbot: Mixed Methods Study. *JMIR Med. Educ.* **2024**, *10*, e51308. [CrossRef] [PubMed] [PubMed Central]
27. Abou-Abdallah, M.; Dar, T.; Mahmudzade, Y.; Michaels, J.; Talwar, R.; Tornari, C. The quality and readability of patient information provided by ChatGPT: Can AI reliably explain common ENT operations? *Eur. Arch Otorhinolaryngol.* **2024**, *281*, 6147–6153. [CrossRef] [PubMed]
28. Ömür Arça, D.; Erdemir, İ.; Kara, F.; Shermatov, N.; Odacioğlu, M.; İbişoğlu, E.; Hanci, F.B.; Sağiroğlu, G.; Hanci, V. Assessing the readability, reliability, and quality of artificial intelligence chatbot responses to the 100 most searched queries about cardiopulmonary resuscitation: An observational study. *Medicine* **2024**, *103*, e38352. [CrossRef] [PubMed]