*Article*

# The Emerging Role of Large Language Models in Improving Prostate Cancer Literacy

**Marius Geantă [1,4]* , Daniel Bădescu [1,2] , Narcis Chirca [1,2], Ovidiu Cătălin Nechita [2], Cosmin George Radu [2] , Ștefan Rascu [1,2], Daniel Rădăvoi [1], Cristian Sima [2], Cristian Toma [1,2], Viorel Jinga [1, 3]**

[1] Department of Urology, "Carol Davila" University of Medicine and Pharmacy, 8 Eroii Sanitari Blvd., 050474 Bucharest, Romania. e-mail: rectorat@umfcd.ro

[2] Department of Urology, "Prof. Dr. Th. Burghele" Clinical Hospital, 20 Panduri Str., 050659 Bucharest, Romania. e-mail: contact@burghele.ro

[3] Academy of Romanian Scientists, 3 Ilfov, 050085, Bucharest, Romania. e-mail: secretariat@aosr.ro

[4] Center for Innovation in Medicine, 42J Theodor Pallady Bvd., 032266, Bucharest, Romania. e-mail: marius.geanta@ino-med.ro

* Correspondence: Marius Geantă, e-mail@marius.geanta@ino-med.ro; Tel.: +40.745.020.878

**Abstract: (1) Background**: Generative AI has the potential to revolutionize patient education and health literacy, particularly through chatbots that provide real-time, personalized health information. **Research Question**: How effective are three widely available Large Language Models (LLMs) - ChatGPT 3.5, Co-Pilot, and Gemini - compared to the official Patient's Guide in delivering accurate, timely, complete, and easily to understand information about prostate cancer? (2) **Methods**: The methodology for this study was designed to systematically assess the efficacy of three LLMs compared to the official Patient's Guide. We formulated 25 representative questions about prostate cancer, validated by clinical experts, and analyzed the responses using a Likert scale that assessed accuracy, timeliness, completeness, and understandability. Appropriate statistical techniques were employed to evaluate the outcomes. (3) **Results**: ChatGPT consistently emerged as the most effective source, receiving high ratings across various criteria, indicating its robustness and reliability as a source of information on prostate cancer. Co-Pilot was also favorably viewed, although its impact was slightly less pronounced than that of ChatGPT. (4) **Conclusions**: While the Guide sets a high standard, the additional benefits provided by ChatGPT and Co-Pilot underscore the importance of continuous improvement and innovation in educational tools, especially in critical health information domains like prostate cancer. The upcoming EU AI Act underscores the necessity for ethical and rigorous oversight in health-related AI applications. Future studies should explore the potential biases in AI responses and examine their impact on patient outcomes.

**Keywords:** prostate cancer, ChatGPT, Co-Pilot, Gemini, cancer literacy, large language models

## 1. Introduction

The public launch in November 2022 of ChatGPT, a large language model (LLM) chatbot that can write informed and precise texts on various subjects, including health, has garnered the attention of the medical and health research community. Translation services, chatbots for customer service, and content generation were the first applications that primarily incorporated LLMs. However, their potential in the medical field quickly became apparent. By processing medical literature, patient records, and other forms of data, LLMs have assisted in tasks ranging from drafting medical documents to providing tentative diagnostic suggestions and generating patient-specific medical advice.

Integrating LLMs into healthcare is part of a broader trend toward digitalization and personalized medicine. These models not only enhance the efficiency of healthcare providers but also play a crucial role in democratizing medical knowledge, thus potentially transforming patient outcomes worldwide [1]. Their ability to quickly synthesize and relay complex medical information can improve health literacy among the general public, a critical factor in the prevention and management of diseases [2]. Patients and the general public have begun to use LLMs to seek information about various diseases, which can impact prevention and therapeutic conduct, adherence to treatment, and, ultimately, therapeutic outcomes [3]. The level of health literacy correlates with better therapeutic outcomes. [4, 5]

Considering the large-scale availability of LLMs and their potential role in the field of health literacy [6, 7], the study aimed to evaluate critically, from the perspective of cancer literacy, the performance of three large language models—ChatGPT, Gemini, and Co-Pilot—compared to the Patient's Guide on prostate cancer [8 -10]. Four parameters were considered - accuracy, timeliness, comprehensiveness, and easy-to-use - for evaluating the answers provided by the three LLMs and the Patient's Guide to 25 key questions about prostate cancer. Our statistical analysis claims that ChatGPT and Co-Pilot performed better than Gemini and the Patient's Guide in providing answers to prostate cancer.

## 2. Materials and Methods

### *2.1 Study Design and Question Formulation*

The methodology for this study was designed to systematically assess the efficacy of three large language models (LLMs: Co-Pilot, ChatGPT, Gemini) compared to the official Patient's Guide in providing *accurate*, *timeliness*, *comprehensible*, and *easy-to-use* information on prostate cancer. We formulated 25 questions reflecting common queries related to prostate cancer. An English version of the questionnaire (the 25 questions) as well as the raw data frame are freely available. [11]

The following prompt was used to interrogate the three LLMs: *I am a man, and my doctor has informed me that I have been diagnosed with prostate cancer. I am interested in learning more about the diagnosis, treatment, and overall management of the disease, which will help me better manage the condition and improve my quality of life. Therefore, I have the following questions for which I would like to obtain answers*.

For each question, responses were generated using two general sources: the established official Patient's Guide and three advanced LLMs—ChatGPT 3.5, Gemini, and Co-Pilot. A single operator queried all three models to ensure consistency in the data collection process. The queries were conducted using incognito mode in Google Chrome to eliminate any personalized search biases, ensuring that each LLM responded based solely on their built-in knowledge and algorithms.

### *Blinding and Randomization of Responses*

After collecting the responses, we performed a randomization process to mix the answers thoroughly. This procedure was to ensure that the subsequent evaluation by experts would be free from preconceived notions about each response's source.

### *Experts and Expert Evaluation*

The randomized responses were then presented to a panel of eight experts in prostate cancer (i.e., medical doctors). These experts are affiliated with the foremost hospital in Bucharest, Romania, which is noted for treating the largest number of prostate cancer patients annually. We targeted this hospital to ensure that we have access to the most prominent Romanian medical doctors in this medical field, in Romania. We sent invitations to all the medical doctors that treat prostate cancer patients and are affiliated with this hospital of interest. Eventually, we ended up with a convenience sample of eight experts.

All experts are males, have an average age of 38.25 years (*SD*: 7.13, *Range*: 20) and an average number of patients per month of 16.88 (*SD*: 25.84, *Range*: 79). We note that experts display a low to moderate variability in terms of age (*Coefficient of variation:*

18.63%). On the other hand, they exhibit a high variability in terms of cancer patients treated per month (*Coefficient of variation:* 153.11%); this suggests a highly skewed distribution.

The experts were blinded to the source of each response to maintain the integrity of the assessment process. Furthermore, we implemented this process to reduce the disparities and potential prejudices arising from the variation among medical practitioners in terms of age and the number of cancer patients treated. However, it is important to interpret the data cautiously because the sample is homogenous in terms of sex assigned at birth, with all panel members being male medical practitioners. Due to the limited and newly growing research on this subject, there are no previous studies available regarding the influence of assigned sex at birth on the distribution of answers. However, it is possible that there may be biases in the replies related to this socio-demographic aspect, specifically the sex assigned at birth.

Every member of the panel was provided with a digital version of the questionnaire. Subsequently, we pooled all the responses into a data frame and conducted statistical analysis utilizing the R utilities accessible in RStudio. All participants voluntarily agreed to participate in the study after receiving a consent form. This document provides information on the research's objectives and context. It also highlights that participants' identities will be kept anonymous and their involvement will be treated with utmost confidentiality. Furthermore, it emphasizes that participation in the study is entirely voluntary. No incentives, whether monetary or non-monetary, were provided to the research participants. However, we made a commitment to grant them access to the data frame and any scientific documents (such as study reports, scientific publications, oral talks, etc.) that are based on the collected data.

All methods were carried out in accordance with the relevant national and international guidelines and regulations. Informed consent was obtained from all participants. The privacy rights of the study participants were observed.

### *Evaluation Criteria and Scoring*

Each expert independently evaluated the responses based on four key criteria (ATCE algorithm): accuracy, timeliness, comprehensiveness, and easy-to-use. Each criterion was rated on a Likert scale ranging from 1 (poor) to 5 (excellent). This scoring system allowed us to quantitatively assess the quality and utility of the information provided by each source.

### *Language and Cultural Considerations*

The entire evaluation process was conducted in Romanian, which not only facilitated a natural understanding among the native expert panel but also enabled an assessment of how effectively the LLMs could handle and reflect local and cultural nuances in their responses. This approach will inform the future development of ethical, diverse, equitable, and inclusive human-LLM collaborative models to improve literacy concerning prostate cancer.

## *2.2 Statistical analysis*

We implemented a range of statistical techniques that were appropriate for achieving the goals of our study. Specifically, our interest was in determining: a) if the Guide surpasses each of the three LLMs, and b) which information source is the most effective in the context of our study design.

We aggregated the scores assigned by each of the eight experts to each of the 25 questions by sources (tools) for information. Firstly, we performed the aggregations per each criterion (accuracy, timeliness, comprehensiveness, and easy-to-use). Then, we performed a grant aggregation, i.e., we computed the sum of all scores irrespective of the four criteria). We fit linear mixed-effects models by Restricted Maximum Likelihood (REML) to the resulting aggregation of scores. We selected this family of models as we wanted to control for the variations attributable to differences across experts, i.e., observable differences such as age, number of patients, and latent differences. This family of statistical models is useful as allows for separating the fixed effects (the differences among the ratings given by the experts) from the random effects (modeling the dependencies and non-independence among data points due to the grouping structure, i.e., measurements for each information source are nested in experts). We performed the test of the mixed-effects models using the algorithms implemented in the *lme4* R package. Also, we performed pairwise comparisons using *emmeans* R package. Specifically, we were interested in comparing information sources with the purpose to ascertain the performance of each tool in relation to the others (e.g., ChatGPT vs Co-Pilot, ChatGPT vs the Guide, ChatGPT vs Gemini, etc.).

Before running the statistical analysis (fitting the mixed-effects models and performing pairwise comparisons), we checked in our data for different assumptions. Firstly, we tested the assumption of normality of the residuals using the Shapiro-Wilk test (i.e., whether the score distribution for each source of information deviates from a normal distribution). Secondly, we tested for the assumption of the homogeneity of variances. For this purpose, we used Levene's Test (the *car* R package) to understand if the scores are equal across different levels of the sources of information. Additionally, we used other two similar tests: the Bartlett's test (this tends to be more robust when data are normally distributed) and the Fligner-Killeen Test (this is less sensitive to the normality of distributions). We used the Levene's Test, the Bartlett's test and the Fligner-Killeen Test to reach a more comprehensive overview of variance homogeneity and to provide solid grounds for parametric test application. Each medical specialist gave complete responses to the 25 queries which resulted in no missing data. For replication purposes, the code and the data are freely available [11] .

## 3. Results

In **Table 1**, we report the aggregated distribution of ratings (grades or scores) that panel members gave to each information source. We provide the distribution of total scores per each assessment criteria (accuracy, timeliness, comprehensiveness, easy-to-use) and the grant sum of scores (accounting for all the criteria).

**Table 1.** Distributions of aggregated scores given by panel experts.

| Criterion | experts' id | ChatGPT | Gemini | Co-Pilot | Guide |
|---|---|---|---|---|---|
| *Grant total* | | | | | |
| | 1 | 432 | 373 | 432 | 409 |
| | 2 | 408 | 275 | 335 | 249 |
| | 3 | 377 | 341 | 377 | 343 |
| | 4 | 434 | 394 | 410 | 376 |
| | 5 | 411 | 388 | 394 | 372 |
| | 6 | 456 | 363 | 416 | 359 |
| | 7 | 451 | 406 | 448 | 435 |
| | 8 | 363 | 354 | 366 | 349 |
| *Accuracy* | | | | | |
| | 1 | 109 | 93 | 105 | 101 |
| | 2 | 99 | 66 | 78 | 62 |

| | | | | | |
|---|---|---|---|---|---|
| | 3 | 96 | 80 | 93 | 81 |
| | 4 | 112 | 101 | 101 | 101 |
| | 5 | 100 | 96 | 95 | 88 |
| | 6 | 113 | 91 | 106 | 90 |
| | 7 | 109 | 96 | 106 | 102 |
| | 8 | 88 | 79 | 84 | 77 |
| *Timeliness* | | | | | |
| | 1 | 108 | 93 | 110 | 104 |
| | 2 | 100 | 66 | 77 | 66 |
| | 3 | 98 | 80 | 101 | 98 |
| | 4 | 112 | 101 | 103 | 104 |
| | 5 | 99 | 96 | 93 | 94 |
| | 6 | 109 | 91 | 100 | 95 |
| | 7 | 112 | 96 | 115 | 114 |
| | 8 | 92 | 79 | 97 | 98 |
| *Comprehensiveness* | | | | | |
| | 1 | 98 | 78 | 102 | 90 |
| | 2 | 101 | 63 | 86 | 58 |
| | 3 | 94 | 74 | 90 | 78 |
| | 4 | 106 | 87 | 96 | 80 |
| | 5 | 97 | 89 | 93 | 82 |
| | 6 | 118 | 79 | 106 | 78 |
| | 7 | 109 | 89 | 107 | 99 |
| | 8 | 93 | 86 | 89 | 87 |
| *Easy-to-use* | | | | | |
| | 1 | 117 | 108 | 115 | 114 |
| | 2 | 108 | 75 | 94 | 63 |
| | 3 | 89 | 93 | 93 | 86 |
| | 4 | 104 | 105 | 110 | 91 |
| | 5 | 115 | 109 | 113 | 108 |

|  |  |  |  |  |
|---|---|---|---|---|
| 6 | 116 | 100 | 104 | 96 |
| 7 | 121 | 113 | 120 | 120 |
| 8 | 90 | 95 | 96 | 87 |

Note. We aggregated experts' ratings (1 to 5) across all 25 questions. The *Grant total* is computed over all the aggregations performed across the four assessment criteria: *accuracy, timeliness, comprehensiveness,* and *easy-to-use*.

**Table 2** illustrates the results of five linear mixed-effects models that fit the data structure presented in **Table 1**. We fit these models to understand how the panel experts rated the four specific sources (i.e., ChatGPT, Co-Pilot, Gemini, and the Guide) and their effectiveness in providing information related to prostate cancer. The Guide (or the assessments associated with the information conveyed by the Guide) stands as the baseline in all the models reported in **Table 2**.

We fit Model 1 on the total scores elicited by the panel experts (this corresponds to the *Grant total* in **Table 1**). According to this model, there is significant variation among experts, indicating differing baseline opinions. However, the intercept (Est. 361.50, *p* < .000) shows a high score that, on its own, indicates the Guide to be a pretty effective information source. In other words, this sets a high standard for the other sources or tools. ChatGPT (Est. 55.00, *p* < .001) and Co-Pilot (Est. 35.75, *p* < .01) provide statistically significant improvements over the Guide, indicating their additional benefits. At the same time, Gemini does not significantly alter the perception (Est. = 0.25, *p* = .98), suggesting it offers no improvement over the Guide.

**Table 2.** Linear mixed-effects models fit by REML

|  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|
| **Model 1: General** |  |  |  |  |  |  |
| *Random effects* | Variance | Std. Dev. |  |  |  |  |
| Groups (intercept) | 1266.8 | 35.59 |  |  |  |  |
| Residual | 531.8 | 23.06 |  |  |  |  |
| *Fixed intercepts* |  |  |  |  |  |  |
|  | Estimate | SE | df | t value | Pr(>ltl) |  |
| (Intercept) | 361.50 | 14.99 | 11.25 | 24109 | 0.000000 | *** |
| Co-Pilot | 35.75 | 11.53 | 21.00 | 3100 | 0.005418 | ** |
| Gemini | 0.25 | 11.53 | 21.00 | 0.022 | 0.982907 |  |
| ChatGPT | 55.00 | 11.53 | 21.00 | 4770 | 0.000103 | *** |
| **Model 2: Accuracy** |  |  |  |  |  |  |
| *Random effects* | Variance | Std. Dev. |  |  |  |  |
| Groups (intercept) | 105.11 | 10.252 |  |  |  |  |
| Residual | 26.27 | 5.125 |  |  |  |  |
| *Fixed intercepts* |  |  |  |  |  |  |
|  | Estimate | SE | df | t value | Pr(>ltl) |  |
| (Intercept) | 87.80 | 4.05 | 9.59 | 21654 | 21.00 | *** |

|  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|
| Co-Pilot | 8.25 | 2.56 | 21.00 | 3219 | 0.00411 | ** |
| Gemini | 0.00 | 2.56 | 21.00 | 0.000 | 100000 |  |
| ChatGPT | 15.50 | 2.56 | 21.00 | 6049 | 0.00000 | *** |
| **Model 3: Timeliness** |  |  |  |  |  |  |
| *Random effects* | Variance | Std. Dev. |  |  |  |  |
| Groups (intercept) | 88.73 | 9.419 |  |  |  |  |
| Residual | 40.95 | 6.399 |  |  |  |  |
| *Fixed intercepts* |  |  |  |  |  |  |
|  | Estimate | SE | df | t value | Pr(>\|t\|) |  |
| (Intercept) | 96625 | 11645 | 24000 | 21654 | 0.0000 | *** |
| Co-Pilot | 2875 | 21000 | 0.899 | 3219 | 0.3791 |  |
| Gemini | -8875 | 21000 | -2774 | 0.000 | 0.0114 | * |
| ChatGPT | 7125 | 21000 | 2227 | 6049 | 0.0370 | * |
| **Model 4: Comprehensiveness** |  |  |  |  |  |  |
| *Random effects* | Variance | Std. Dev. |  |  |  |  |
| Groups (intercept) | 39.37 | 6.274 |  |  |  |  |
| Residual | 50.66 | 7.118 |  |  |  |  |
| *Fixed intercepts* |  |  |  |  |  |  |
|  | Estimate | SE | df | t value | Pr(>\|t\|) |  |
| (Intercept) | 81500 | 3355 | 17793 | 24295 | 0.0000 | *** |
| Co-Pilot | 14625 | 3559 | 21000 | 4110 | 0.0005 | *** |
| Gemini | -0.875 | 3559 | 21000 | -0.246 | 0.8082 |  |
| ChatGPT | 20500 | 3559 | 21000 | 5760 | 0.0000 | *** |
| **Model 5: Easy-to-use** |  |  |  |  |  |  |
| *Random effects* | Variance | Std. Dev. |  |  |  |  |
| Groups (intercept) | 132.68 | 11.518 |  |  |  |  |
| Residual | 52.87 | 7.271 |  |  |  |  |
| *Fixed intercepts* |  |  |  |  |  |  |
|  | Estimate | SE | df | t value | Pr(>\|t\|) |  |
| (Intercept) | 95625 | 4816 | 11050 | 19856 | 0.00000 | *** |
| Co-Pilot | 10000 | 3636 | 21000 | 2751 | 0.01198 | * |

| | | | | | | |
|---|---|---|---|---|---|---|
| Gemini | 4125 | 3636 | 21000 | 1135 | 0.26932 | |
| ChatGPT | 11875 | 3636 | 21000 | 3266 | 0.00369 | ** |

Note. In each model, we have 32 observations and eight experts. The t-tests use Satterthwaite's method. The Guide is the baseline in each model.

The results corresponding to Model 2 reveal a lower variance than Model 1, indicating more consistency in expert opinions for the *Accuracy* criterion. ChatGPT (Est. 15.50, *p* < .000) and Co-Pilot (Est. 8.25, *p* < .01) are valuable in terms of the accuracy of the information provision. Again, ChatGPT is particularly influential for this specific criterion. Model 3 exhibits variability among experts concerning the *timeliness* of the responses generated by the four sources of information. Furthermore, ChatGPT provides a consistent improvement compared to the baseline (Est. 7125, *p* < .05).

Model 4 displays the lowest variability, indicating strong consensus among experts regarding the *comprehensiveness* dimension of the responses. CO-Pilot (Est. 14625, *p* < .001) and ChatGPT (Est. 20500, *p* < .000) are seen as highly effective, with ChatGPT showing the most substantial positive effect. Model 5 indicates a moderate consensus among experts concerning the *easy-to-use* evaluation dimension. As in the previous models, ChatGPT (Est. 11875, *p* < .01) and Co-Pilot (Est. 10000, *p* < .05) enhance ratings significantly.

As a general commentary, ChatGPT consistently emerges as the most effective source across different criteria, receiving high ratings from panel experts. This suggests its robustness and reliability as a source of prostate cancer information. Co-Pilot is favorably viewed, though its impact is slightly less pronounced than ChatGPT. However, experts still consider it a valuable tool. Gemini is viewed either neutrally or negatively across models. This variability suggests that while it may have uses, it might not be the best source for disseminating prostate cancer information.

In all the models, the Guide (as baseline) remains consistently high, suggesting it is a robust tool across various specific criteria. While the Guide is practical, ChatGPT and Co-Pilot introduce additional features or present information in a way that the experts find even more helpful or accessible. Gemini presents a non-significant effect (except for Model 2, where it is negative) that suggests it does not consistently offer improvements over the Guide.

**Table 3** reports a series of pair-wise comparisons between the information tools that the experts evaluated. These post hoc tests are necessary to indicate which tools differ from each other and how. As indicated in **Table 3**, we associate these post hoc tests with the linear mixed-effects models reported in **Table 2**.

Based on the information available in **Table 3**, we state that, across all models, ChatGPT consistently emerges as the most effective tool, often showing significant improvements over the Guide and other tools. Co-Pilot performs better than the Guide and is comparable to other tools but does not consistently surpass ChatGPT. Gemini shows the least consistent performance, often not significantly better than the Guide, and is usually less effective than Co-Pilot and ChatGPT.

For instance, the post hoc tests corresponding to Model 1 illustrate that ChatGPT is significantly more effective than the Guide (Est. -55.00, *p* < .001) and then the Gemini (Est. -54.75, *p* < .001). Even if there is no significant difference between ChatGPT and Co-Pilot (Est. -19.35, *p* = 0.36), the numerical difference marks a slight preference for ChatGPT among the experts.

**Table 3.** Post hoc tests for comparing sources of information

| contrast | estimate | SE | df | t.ratio | p.value |
|---|---|---|---|---|---|
| ***Model 1: General*** | | | | | |
| Guide-copilot | -35.75 | 11.05 | 21 | -3.100 | 0.0257 |
| Guide-Gemini | -0.25 | 11.05 | 21 | -0.022 | 1.0000 |
| Guide-ChatGPT | -55.00 | 11.05 | 21 | -4.770 | 0.0006 |
| Copilot-Gemini | 35.50 | 11.05 | 21 | 3.079 | 0.0270 |

|  |  |  |  |  |  |
|---|---|---|---|---|---|
| Copilot-ChatGPT | -19.25 | 11.05 | 21 | -1.669 | 0.3638 |
| Gemini-ChatGPT | -54.75 | 11.05 | 21 | -4.748 | 0.0006 |
| ***Model 2: Accuracy*** |  |  |  |  |  |
| Guide-Copilot | -8.25 | 2.56 | 21 | -3.219 | 0.0198 |
| Guide-Gemini | 0.00 | 2.56 | 21 | 0.000 | 1.0000 |
| Guide-ChatGPT | -15.50 | 2.56 | 21 | -6.049 | <.0001 |
| Copilot-Gemini | 8.25 | 2.56 | 21 | 3.219 | 0.0198 |
| Copilot-ChatGPT | -7.25 | 2.56 | 21 | -2.829 | 0.0458 |
| Gemini-ChatGPT | -15.50 | 2.56 | 21 | -6.049 | <.0001 |
| ***Model 3: Timeliness*** |  |  |  |  |  |
| Guide-Copilot | -2.88 | 3.2 | 21 | -0.899 | 0.8057 |
| Guide-Gemini | 8.88 | 3.2 | 21 | 2.774 | 0.0514 |
| Guide-ChatGPT | -7.12 | 3.2 | 21 | -2.227 | 0.1485 |
| Copilot-Gemini | 11.75 | 3.2 | 21 | 3.672 | 0.0072 |
| Copilot- ChatGPT | -4.25 | 3.2 | 21 | -1.328 | 0.5559 |
| Gemini- ChatGPT | -16.00 | 3.2 | 21 | -5.001 | 0.0003 |
| ***Model 4: Comprehensiveness*** |  |  |  |  |  |
| Guide-Copilot | -14.625 | 3.56 | 21 | -4.110 | 0.0026 |
| Guide-Gemini | 0.875 | 3.56 | 21 | 0.246 | 0.9946 |
| Guide- ChatGPT | -20.500 | 3.56 | 21 | -5.760 | 0.0001 |
| Copilot-Gemini | 15.500 | 3.56 | 21 | 4.355 | 0.0015 |
| Copilot-ChatGPT | -5.875 | 3.56 | 21 | -1.651 | 0.3734 |
| Gemini-ChatGPT | -21.375 | 3.56 | 21 | -6.006 | <.0001 |
| ***Model 5: Easy-to-use*** |  |  |  |  |  |
| Guide-Copilot | -10.00 | 3.64 | 21 | -2.751 | 0.0539 |
| Guide-Gemini | -4.12 | 3.64 | 21 | -1.135 | 0.6729 |

| | | | | | |
|---|---|---|---|---|---|
| Guide-ChatGPT | -11.88 | 3.64 | 21 | -3.266 | 0.0179 |
| Copilot-Gemini | 5.88 | 3.64 | 21 | 1.616 | 0.3915 |
| Copilot-ChatGPT | -1.88 | 3.64 | 21 | -0.516 | 0.9544 |
| Gemini-ChatGPT | -7.75 | 3.64 | 21 | -2.132 | 0.1757 |

**Table 4** reports the results of the tests used to assess the normality of data distributions and the homogeneity of variances across the different groups represented by each source of information. We used Shapiro-Wilk test (W, the assumption of normality), Levene's test, Bartlett's K-squared and Fligner-Killeen Test (homogeneity of variance). All tools (sources of information) show p-values well above 0.05, suggesting that the scores are normally distributed for each tool. The consistent results across Levene's, Bartlett's, and Fligner-Killeen tests indicate that the assumption of equal variances holds true for all categories. These diagnostics support the use of linear mixed-effects models that were reported in **Table 2**.

**Table 4.** Tests for normality and homogeneity of variances

| *Normality assumption* | *Homogeneitey of variances* |
|---|---|
| *Overall evaluation* | |
| Guide, W = 0.91237, p = 0.3711 | Levene's test: F(3, 28) = 0.1927, p = 0.9005 |
| Co-Pilot, W = 0.98284, p = 0.9756 | Bartlett's K-squared = 1.945, df = 3, p = 0.5839 |
| Gemini, W = 0.88878, p = 0.2280 | Fligner-Killeen Test:<br>med chi-squared = 0.28729, df = 3, p = 0.9624 |
| ChatGPT, W = 0.93274, p = 0.5414 | |
| *Accuracy evaluation* | |
| Guide, W = 0.90624, p = 0.3284 | Levene's test: F(3, 28) = 0.3158, p = 0.8138 |
| Co-Pilot, W = 0.88498, p = 0.2100 | Bartlett's K-squared = 1.4537, df = 3, p = 0.693 |
| Gemini, W = 0.90508, p = 0.3207 | Fligner-Killeen Test:<br>med chi-squared = 0.97968, df = 3, p = 0.8062 |
| ChatGPT, W = 0.91362, p = 0.3802 | |
| *Timeliness evaluation* | |
| Guide, W = 0.83798, p = 0.0718 | Levene's test: F(3, 28) = 0.1206, p = 0.9472 |
| Co-Pilot, W = 0.94442, p = 0.6551 | Bartlett's K-squared = 2.4542, df = 3, p = 0.4836 |
| Gemini, W = 0.90508, p = 0.3207 | Fligner-Killeen Test:<br>med chi-squared = 0.19041, df = 3, p = 0.9791 |
| ChatGPT, W = 0.90091, p = 0.2944 | |
| *Comprehensiveness* | |
| Guide, W = 0.92876, p = 0.5048 | Levene's test: F(3, 28) = 0.1023, p = 0.9580 |

| | |
|---|---|
| Co-Pilot, W = 0.91887, p = 0.4207 | Bartlett's K-squared = 1.2837, df = 3, p = 0.733 |
| Gemini, W = 0.87805, p = 0.1804 | Fligner-Killeen Test:<br>med chi-squared = 0.090494, df = 3, p = 0.993 |
| ChatGPT, W = 0.91983, p = 0.4285 | |
| *Easy-to-use* | |
| Guide, W = 0.95716, p = 0.7827 | Levene's test: F(3, 28) = 0.675, p = 0.5746 |
| Co-Pilot, W = 0.91231, p = 0.3706 | Bartlett's K-squared = 2.4896, df = 3, p = 0.4772 |
| Gemini, W = 0.90322, p = 0.3088 | Fligner-Killeen Test:<br>med chi-squared = 1.4162, df = 3, p = 0.7017 |
| ChatGPT, W = 0.87014, p = 0.1512 | |

Note. We used Shapiro-Wilk test (W) for checking the normality assumption.

## 4. Discussion

This study's exploration of Large Language Models (LLMs) such as ChatGPT, Gemini, and Co-Pilot has yielded significant insights into their potential to enhance cancer literacy, particularly within prostate cancer and specific cultural contexts. The findings reveal varying degrees of effectiveness among these models in improving prostate cancer information and literacy among patients.

Among the three LLMs evaluated, ChatGPT and Co-Pilot performed better than the third LLM, Gemini, and outperformed the traditional Patient's Guide across all evaluated criteria. Statistically significant differences between ChatGPT and Co-Pilot were not observed, indicating comparable performance levels between these two models. The results are aligned with previous data on the efficacy of the LLMs ChatGPT and Co-Pilot (formerly Bard) in providing accurate, timely, complete, and easily to understand information about prostate cancer [12].

The results underscore the potential of LLMs to enhance the effectiveness of patient and caregiver education regarding prostate cancer. The study demonstrates that, for prostate cancer, there are statistically significant differences between the LLMs, with ChatGPT and Co-Pilot emerging as superior sources of LLM-based information. Concurrently, ChatGPT and Co-Pilot are identified as prime candidates for developing personalized virtual assistants [13] to aid patients diagnosed with prostate cancer and their families.

Traditional patient and family education methods [14] like the Patient's Guide could also benefit from developing LLMs. In the future, LLMs could contribute to creating dynamic guides that offer higher accuracy, more current and consistent information, and are more accessible for patients and their families to understand, co-created by physicians and patients. [15 - 16]

It is acknowledged that using LLMs raises ethical questions [17], particularly concerning the accuracy of machine-generated advice and its impact on patient decision-making. The role of physicians [18] is essential in ensuring the reliability of these tools and establishing clear guidelines for their use to prevent misinformation and ensure the quality of information delivered to patients and families. For these reasons, the development of a human-LLM collaborative model is crucial [19]. In the AI era, the traditional linear model of physician-patient communication [20] is transforming into a complex and dynamic model [21] where the professional authority (the physician) must actively and continuously contribute to developing, training, and refining LLM-based chatbots.

At the same time, the beneficiary (the patient and family) evolves from a passive recipient of information into an active contributor.

*Future Directions*

There is immense potential for integrating LLMs more deeply into the healthcare systems. Developing models that can interact seamlessly with electronic health records (EHRs) to provide contextual advice could revolutionize patient care. [22 - 24] Additionally, further research should focus on personalizing LLMs interactions based on individual patient histories to enhance the relevance and effectiveness of the information provided. This underscores the need for regulatory frameworks to oversee the deployment of LLMs in healthcare settings. [25] Such regulations should ensure these tools meet stringent accuracy and safety standards like other medical devices. The conclusions of the study resonate with the recently approved EU AI Act [26] who will be effective from 2026, a key document highlighting the need for expert oversight of the high-risk AI systems such as the LLMs used in the health contexts.

Our findings suggest that the Guide is a solid foundation for providing information about prostate cancer. However, ChatGPT and Co-Pilot present enhancements that recommend their incorporation in information dissemination strategies, possibly making the information more engaging, accessible, or comprehensible. Decisions about which tool to use or recommend should consider these differences in effectiveness. Tools that significantly improve the Guide could be prioritized for situations requiring higher engagement or more profound understanding. Understanding that Gemini does not improve upon the Guide might lead to reconsidering its use or pushing for its development to meet the guidelines and other tools.

In summary, while the Guide sets a high standard of effectiveness, the additional benefits provided by ChatGPT and Co-Pilot underline the importance of continuous improvement and innovation in educational tools, especially in critical health information domains like prostate cancer.

Our results can guide healthcare providers, researchers, and decision-makers in optimizing the tools and resources they deploy for education and communication about prostate cancer, ensuring that the most effective platforms are utilized to disseminate crucial health information.

## 5. Conclusions

As these models continue to evolve, their influence on the medical field is expected to grow, making their study and understanding an essential area of research. The use of LLMs like ChatGPT and Co-Pilot in improving cancer literacy among prostate cancer patients holds promising potential. However, continuous improvements, rigorous testing, and thoughtful integration into clinical practice, accompanied by appropriate ethical and regulatory oversight, are essential to fully realize their benefits without compromising patient safety or quality of care.

## 6. Patents

Not applicable.

**Supplementary Materials:** The data, the questionnaire, the code and other related files are freely available for replication and secondary data analysis at: https://doi.org/10.5281/zenodo.11217682

**Author Contributions:** Conceptualization, Marius Geantă and Viorel Jinga; Formal analysis, Marius Geantă; Methodology, Marius Geantă and Viorel Jinga; Supervision, Viorel Jinga; Validation, Daniel Bădescu, Narcis Chirca, Ovidiu Nechita, Cosmin Radu, Stefan Rascu, Daniel Rădăvoi, Cristian Sima and Cristian Toma; Writing – original draft, Marius Geantă; Writing – review & editing, Marius Geantă and Viorel Jinga. 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:** Informed consent was obtained from all subjects involved in the study.

**Data Availability Statement:** The data, the questionnaire, the code and other related files are freely available for replication and secondary data analysis at https://doi.org/10.5281/zenodo.11217682

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

## References

1. Topol, E.J. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med* **25**, 44–56 (2019). https://doi.org/10.1038/s41591-018-0300-7

2. Clusmann, J., Kolbinger, F.R., Muti, H.S. *et al.* The future landscape of large language models in medicine. *Commun Med* **3**, 141 (2023). https://doi.org/10.1038/s43856-023-00370-1

3. Haupt CE, Marks M. AI-Generated Medical Advice—GPT and Beyond. *JAMA.* 2023;329(16):1349–1350. doi:10.1001/jama.2023.5321

4. Walters, R., Leslie, S.J., Polson, R. *et al.* Establishing the efficacy of interventions to improve health literacy and health behaviours: a systematic review. *BMC Public Health* **20**, 1040 (2020). https://doi.org/10.1186/s12889-020-08991-0

5. Shahid, R., Shoker, M., Chu, L.M. *et al.* Impact of low health literacy on patients' health outcomes: a multicenter cohort study. *BMC Health Serv Res* **22**, 1148 (2022). https://doi.org/10.1186/s12913-022-08527-9

6. Amin KS, Mayes LC, Khosla P, Doshi RH. Assessing the Efficacy of Large Language Models in Health Literacy: A Comprehensive Cross-Sectional Study. Yale J Biol Med. 2024 Mar 29;97(1):17-27. doi: 10.59249/ZTOZ1966. PMID: 38559461; PMCID: PMC10964816.

7. Miriam McMullan. Patients using the Internet to obtain health information: How this affects the patient–health professional relationship,Patient Education and Counseling, Volume 63, Issues 1–2, 2006,Pages 24-28,
ISSN 0738-3991, https://doi.org/10.1016/j.pec.2005.10.006.

8. Federatia Asociatiilor Bolnavilor de Cancer. Available online: https://cancerprostata.fabc.ro/wp-content/uploads/2023/02/Mic_Ghid_de_diagnostic_si_tratament_pentru_pacient_Cancerul-de-prostata.pdf (Accessed 12 May 2024)

9. 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* **21**, 269 (2023). https://doi.org/10.1186/s12967-023-04123-5

10. Iannantuono GM, Bracken-Clarke D, Floudas CS, Roselli M, Gulley JL, Karzai F. Applications of large language models in cancer care: current evidence and future perspectives. Front Oncol. 2023 Sep 4;13:1268915. doi: 10.3389/fonc.2023.1268915. PMID: 37731643; PMCID: PMC10507617.

11. Geantă M. Large Language Models and prostate cancer. [Data set]. Zenodo. 2024. https://doi.org/10.5281/zenodo.11217682

12. Ahmed Alasker, Seham Alsalamah, Nada Alshathri et al. Performance of Large Language Models (LLMs) in Providing Prostate Cancer Information, 31 October 2023, PREPRINT (Version 1) available at Research Square [https://doi.org/10.21203/rs.3.rs-3499451/v1]

13. Sezgin E. Redefining Virtual Assistants in Health Care: The Future With Large Language Models
J Med Internet Res 2024;26:e53225. URL: https://www.jmir.org/2024/1/e53225. DOI: 10.2196/53225

14. Marcus C. Strategies for improving the quality of verbal patient and family education: a review of the literature and creation of the EDUCATE model. Health Psychol Behav Med. 2014 Jan 1;2(1):482-495. doi: 10.1080/21642850.2014.900450.

15. Abd-Alrazaq A, AlSaad R, Alhuwail D, Ahmed A, Healy PM, Latifi S, Aziz S, Damseh R, Alabed Alrazak S, Sheikh J. Large Language Models in Medical Education: Opportunities, Challenges, and Future Directions. JMIR Med Educ. 2023 Jun 1;9:e48291. doi: 10.2196/48291

16. Lucas HC, Upperman JS, Robinson JR.A systematic review of large language models and theirimplications in medical education.Med Educ. 2024;1-10.doi:10.1111/medu.1540210LUCASET AL.

17. Li, Hanzhou et al. Ethics of large language models in medicine and medical research. The Lancet Digital Health, Volume 5, Issue 6, e333 - e335. https://doi.org/10.1016/S2589-7500(23)00083-3

18. Uriel K. et al. GPT versus Resident Physicians — A Benchmark Based on Official Board Scores. NEJM AI 2024; 1(5). doi: 10.1056/AIdbp2300192

19. Bano, M., Zowghi, D., & Whittle, J. (2023). AI and Human Reasoning: Qualitative Research in the Age of Large Language Models. *The AI Ethics Journal*, *3*(1). https://doi.org/10.47289/AIEJ20240122

20. L.M.L. Ong, J.C.J.M. de Haes, A.M. Hoos, F.B. Lammes, Doctor-patient communication: A review of the literature, Social Science & Medicine, ISSN 0277-9536, https://doi.org/10.1016/0277-9536(94)00155-M.

21. Chen, Shan et al. The effect of using a large language model to respond to patient messages. The Lancet Digital Health. doi: https://doi.org/10.1016/S2589-7500(23)00083-3

22. Guevara, M., Chen, S., Thomas, S. *et al.* Large language models to identify social determinants of health in electronic health records. *npj Digit. Med.* **7**, 6 (2024). https://doi.org/10.1038/s41746-023-00970-0

23. Lerner, J., Tranmer, M., Mowbray, J., Hâncean, M.-G. REM beyond dyads: relational hyperevent models for multi-actor interaction networks. arXiv:1912.07403. https://doi.org/10.48550/arXiv.1912.07403
24. Lerner J, Hâncean M-G. Micro-level network dynamics of scientific collaboration and impact: Relational hyperevent models for the analysis of coauthor networks. Network Science. 2023;11(1):5-35. doi:10.1017/nws.2022.29
25. Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. NPJ Digit Med. 2023 Jul 6;6(1):120. doi: 10.1038/s41746-023-00873-0.
26. European Parliament. Available online: https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence (accessed on 14 May 2024)