Optimizing ChatGPT: Applying prompt engineering to frequently asked questions in urologic oncology

Mark N. Alshak, Michelle I. Higgins, Craig Cronin, William S. Azar, Joseph G. Cheaib, Max Kates, Sunil H. Patel
The James Buchanan Brady Urological Institute, Johns Hopkins University School of Medicine, Baltimore, MD, United States

Cite as: Alshak MN, Higgins MI, Cronin C, et al. Optimizing ChatGPT: Applying prompt engineering to frequently asked questions in urologic oncology. *Can Urol Assoc J* 2026 March 30; Epub ahead of print. http://dx.doi.org/10.5489/cuaj.9578

Published online March 30, 2026

Corresponding author: Dr. Mark N. Alshak, The James Buchanan Brady Urological Institute, Johns Hopkins University School of Medicine, Baltimore, MD, United States; Malshak1@jh.edu

***

ABSTRACT

Introduction: Patients rely on online searches for patient education materials (PEMs). PEMs are recommended to be written at or below a sixth-grade reading level but are regularly written at a college reading level. Using prompt engineering, we assess the information, misinformation, and readability of ChatGPT responses to urologic oncology questions.

KEY MESSAGES

- ChatGPT provides highly accurate urologic oncology information, but baseline responses are written at an advanced reading level and lack actionable guidance, explanations of risks/benefits, visual aids, and citations.
- Prompt engineering dramatically improves response quality without increasing misinformation.
- Urologists should understand and apply prompt engineering strategies for optimized AI-generated patient education.

Methods: Forty-five questions relating to prostate, bladder, and kidney cancer were presented to ChatGPT (version 4o, OpenAI). Quality of health information was assessed using DISCERN (1 [low] to 5 [high]). Understandability and actionability were assessed using PEMAT-P (0 [low] - 100% [high]). Misinformation was scored from 1 [no misinformation] to 5 [high misinformation]. Grade and reading level were calculated using the Flesch-Kincaid scale [5 (easy) to 16 (difficult), and 100-90 (5th grade level) to 10-0 (professional level), respectively]. Prompt engineering was then applied to responses and evaluated.

Results: ChatGPT answers are highly accurate but too advanced of a reading level and lacked explanations of benefits, risks, visual aids, actionability, and citations. Using prompt engineering,

© 2026 Canadian Urological Association

DISCERN (3.42–4.47, p<0.0001), PEMAT-P understandability (88.4–95.5%, p<0.0001), and actionability (25.6–84.2%, p<0.0001), grade reading level (10.5–5.3, p<0.0001), and reading level (42 [college level] to 71.7 [7th grade], p<0.0001), all significantly improved. Misinformation did not change significantly.
Conclusions: Using prompt engineering, ChatGPT provides highly accurate and understandable PEMs at a patient appropriate reading level and provides concrete resources for patient action. Urologists should understand prompt engineering and be involved in the development of artificial chatbots to optimize results.

INTRODUCTION

Health literacy is the ability of a patient to understand and contribute to their own healthcare and is crucial in optimizing health outcomes.[1] Low health literacy contributes to worse clinical outcomes for patients, including increased emergency department visits, readmissions, and complications following oncologic surgery.[2,3] This holds true in urologic oncology as patients with low health literacy have worse outcomes both before and after surgery, including lower rates of follow-up and increased complications post-operatively.[4]

To address this major issue, patient education materials (PEMs) have been developed to provide patients with the knowledge they need to adequately inform themselves in the peri-operative period. The National Institutes of Health recommends PEMs to be at a reading level at or below the sixth grade as the average United States adult reads at an 8th grade level and the average Medicaid enrollee reads at a 5th grade level.[5,6] However, PEMs in urology are consistently written at much higher levels.[7-9]

An emerging tool that is being investigated is the role of Large Language Models and natural language processing using Artificial Intelligence to create chatbots capable of providing PEMs. The most well-known is chat generative pretrained transformer (ChatGPT), that was developed by OpenAI (San Francisco, CA). Its responses have been shown to be useful and accurate. However, it has shortcomings, such as providing answers at a significantly higher than recommended reading level, and lacking concrete, actionable instructions for patients.[10] Moreover, how a question is prompted to ChatGPT can contribute to performance variability in providing clinically meaningful and accurate urologic information.[11]

Herein we assess the quality of information and readability of ChatGPT responses to urologic oncology frequently asked questions (FAQs). We aim to use prompt engineering to improve the understandability and actionability of responses. We hypothesize that with proper prompt engineering, we can create urologic oncology PEMs that are both highly accurate and understandable to the average patient.

© 2026 Canadian Urological Association

METHODS

Question selection

We gathered a list of FAQs for prostate, kidney, and bladder cancer (Supplemental table 1). These questions were curated by going through the websites of top Urology programs as published on the 2024-2025 U.S. News and World Report and from AUA PEMs. Questions were pulled directly from their FAQ sections. Questions were also created by converting common headers or section titles of informational sections to a question, such as "Prostate cancer" to "What is prostate cancer?" When appropriate, questions were phrased from a patient perspective, such as "How do you diagnose prostate cancer?" to help simulate how a patient would interact with an AI chatbot. Questions were then entered into ChatGPT-4 omni (ChatGPT4o), OpenAI's newest and most advanced model as of November 2024.[12] Responses were collected in an offline document in one session for evaluation to minimize ChatGPT memory bias.

Grading and evaluation of questions

ChatGPT answers were evaluated by two urologists blinded to each other's results. DISCERN is validated for judging the quality of written consumer health information on treatment choices, scored on a scale of 1 (low) to 5 (high).[13] The Patient Education Materials Assessment Tool for Printable Materials (PEMAT-P) is validated to measure understandability and actionability for print and audiovisual patient information, scored on a scale of 0% (low) to 100% (high).[14] The Flesch-Kincaid is a validated metric of understandability of a given document in English.[15] The Flesch-Kincaid grade level presents a score as a U.S. grade level. The Flesch-Kincaid reading level is scored from a scale of 100-90 (5^{th} grade level) to 10-0 (professional level). Misinformation was evaluated using a 5-point Likert scale (range: 1 [none] to 5 [high]). A score of 2 reflected minor misinformation that would not lead to significant change in management or education but failed to include common side effects or complications, 3 was misinformation that recommended an incorrect diagnostic or treatment effect but without major patient harm, and a 4 or 5 was severe misinformation in any domain that could result in patient harm. Median word count of each response was collected and did not include any words appearing in a visual aid.

Prompt engineering

After two reviewers evaluated 45 questions, the results were analyzed to determine areas in which to improve upon ChatGPT4o's responses. Using prompt engineering, which is designing, refining, and implementing various prompts that will guide and change the response AI Chatbots, prompts were engineered to target weak areas of DISCERN, PEMAT, and the Flesch-Kincaid grading scales.[16] This methodology of using prompt engineering to improve ChatGPT responses was developed using FAQs in reconstructive urology and has been previously published and described in detail.[17] Our goal was to optimize and validate this methodology in urologic oncology.

© 2026 Canadian Urological Association

Ultimately, the following prompt was used in front of every question: Provide answers as if talking to a 10-year-old by limiting medical terminology, word length, and sentence length. Keep it simple but be thorough. When discussing treatments, always explain how the treatment works, benefits, risks, and what would happen if no intervention were performed. Provide visual aids, pictures, and tables. Address the patient directly as if you were talking to them. Provide a tangible tool (e.g., menu planners, checklists) to help the patient take action. Summarize key points. Have a section devoted to what sources you used and provide references and links to those references used.

Statistical analysis and ethics approval

All statistical analysis was completed in Microsoft Excel (Redmond, WA: Microsoft Corp). Paired comparisons were conducted using two-tailed Wilcoxon signed-rank tests for medians and paired t-tests for means. Inter-rater reliability between two independent reviewers was quantified using quadratic-weighted Cohen's kappa for DISCERN, PEMAT-P, and misinformation scores. Institutional board review was not obtained due to the public nature of ChatGPT's responses. P value of <0.05 was considered statistically significant.

RESULTS

Discern scores

Median overall DISCERN scores for all 45 questions was 4 (range 1-5). The overall average score was 3.42 ± 1.8 and ranged from 3.38 (bladder cancer) to 3.47 (kidney cancer) (Table 1). ChatGPT omitted sources for information, risks of each treatment, what happens if no treatment is performed, or how various treatments would affect quality of life. However, it did excel at describing various treatment options including benefits of undergoing specific treatment options. Inter-rater reliability across DISCERN domains demonstrated substantial agreement ($\kappa = 0.78–0.91$).

PEMAT-P understandability scores

The overall average PEMAT-P Understandability score was 88% ± 4 and ranged from 86% (bladder cancer) to 92% (prostate cancer), with near perfect inter-rater reliability ($\kappa = 1.00$). ChatGPT used everyday language with active voice. Numbers were presented in a way that was easy to understand and did not require any calculations. Material was presented in an organized manner with headers broken down into shorter sections with a summary provided. However, ChatGPT did not regularly use visual aids, such as graphs or tables, to help present information in an easy-to-understand way.

PEMAT-P actionability scores

The overall average PEMAT-P actionability scores were very low at 26% ± 20 and ranged from 14% (prostate cancer) to 36% (kidney cancer), with near perfect inter-rater reliability ($\kappa = 1.00$). ChatGPT did not always identify an action a patient could take or address the patient directly in

© 2026 Canadian Urological Association

the response. Moreover, it did not provide any tangible tool to help patients take action, such as creation of a checklist or a list of questions that a patient should ask their physician.

Misinformation

ChatGPT's responses were found to be highly accurate with an average score of 1.19 (range 1-3) and ranged from 1.17 (bladder and kidney cancer) to 1.23 (prostate cancer), with fair inter-rater reliability ($\kappa = 0.33$). Of the 45 responses, only 2 (4%) included major misinformation that provided either incorrect treatments, such as diet changes not shown to impact cancer progression or failing to distinguish between fundamentally different surgical treatment options for bladder cancer (Supplementary Table 2). 14 (31%) questions contained minor misinformation that would not otherwise impact a patient's understanding of their disease or treatment options.

Flesch-kincaid readability scores

The overall median Flesch-Kincaid grade level score was very high at 10.5 (6.2-12.7) and ranged from 9.8 (kidney cancer) to 11.1 (bladder cancer). The overall median Flesch-Kincaid reading level was 42, which correlated to a college level, and ranged from 25.4 (college graduate level) to 69.2 (8^{th}-9^{th} grade level). The median word count was 1,058 (range: 254-1,431).

Prompt engineering

Using the above results, it was determined that ChatGPT, while giving accurate answers and good information about disease and treatment options, lacked ability to present responses at a reading level appropriate for the average patient, have more robust discussions about benefits/risks, to provide visual aids and tangible tools, and to provide references/citations for its answers. This methodology has been reported before.[17] Briefly, ChatGPT was prompted specifically to address these deficiencies by trialing different prompts and evaluating the responses.

Using the prompt described in the methods, ChatGPT responses significantly improved across all domains evaluated (Table 2). The average DISCERN score improved significantly from 3.42 to 4.47 ($p<0.0001$). This improvement was driven by ChatGPT citing verified sources and discussing treatments and their associated benefits and risks in much greater detail. The most common references cited were the American Cancer Society and the Mayo Clinic.
PEMAT-P understandability improved significantly (88.4% to 95.5%, $p<0.0001$) driven by ChatGPT using visual aids. PEMAT-P actionability improved dramatically from 25.6% to 84.2% ($p<0.0001$). ChatGPT's responses provided actions and tools for patients to act on those actions, such as checklists or visual aids (Figure 1).

Most importantly, the reading level of ChatGPT's responses were engineered to be of appropriate level for the average patient. At baseline, like most PEMs that are available to patients currently, ChatGPT presented its responses at a very advanced reading level requiring a college-level reading ability. This improved significantly from 42 (college level) to 71.7 (7^{th} grade level), within the recommended 6-8^{th} grade reading level. The Flesch-Kincaid grade level,

© 2026 Canadian Urological Association

another metric of readability, improved from grade 10.5 to 5.3, well within the recommended reading level. Crucially, this improved readability did not result in a significant change in misinformation (1.19 to 1.34, p=0.06) nor did it result in a change in major misinformation (4.4% to 6.7%, p=0.99). Word count was also reduced with the engineered responses from 1,058 to 705 (p<0.0001).

DISUCSSION

We present several important findings in our study. First, we show that ChatGPT provides high quality responses to common patient FAQs in prostate, bladder, and kidney cancer. Cancer patients increasingly engage in online health information seeking information about their specific cancer and treatments. Better informed patients engaging in these activities experience less anxiety about their condition, are more satisfied with their treatments, and are better equipped to participate in their consultations with urologists and oncologists.[18] PEMs provide patient-centric education on their conditions, which helps with informed decision-making, adherence to follow-up, and even improves clinical outcomes.[19] AI chatbots are being increasingly used in Urology to provide patients with information about their disease.[20] Interestingly, previous work has shown that patients prefer AI chatbot responses and find them of higher quality and significantly more empathetic than physician responses.[21] ChatGPT can serve as one of the tools at the disposal of both a patient and a physician to aid patients in navigating a cancer diagnosis.

ChatGPT, while providing high quality and accurate information, did not perform perfectly in all areas evaluated. For example, education level of the PEMs generated prior to prompt engineering far exceeding the recommended reading level, with most at a college or graduate reading level. However, these materials remain written at a level much more advanced than the recommended 6th grader readability level in oncology PEMs in general.[22] In urologic oncology, only 18.6% of PEMs available to patients in the European Association of Urology and America Urology Association meet readability requirements.[8] ChatGPT provides a unique solution to this problem. In our study, we show that with proper prompt engineering, AI chatbots such as ChatGPT can be engineered to provide PEMs at an appropriate reading level. While simplifying medical information raises concerns regarding potential tradeoffs between readability and clinical precision, we found that lowering the reading level did not increase misinformation or meaningfully alter content accuracy. This suggests the nuanced discussions surrounding disease processes, treatments, benefits, risks, and outcomes would not be compromised, and may even be improved if a patient has gained a better understanding of their cancer.

Another area in which ChatGPT did not perform well prior to prompt engineering guidance is actionability. To provide the highest quality of PEMs, education materials need to be presented with actionable content—that is, content that is designed to prompt or suggest action to a patient.[23] ChatGPT excelled in providing highly accurate information about disease processes, but lacked providing tools or actionable content that helps bridge the gap between being informed about a disease and taking steps to address ones disease process. After prompting,

© 2026 Canadian Urological Association

ChatGPT was able to provide excellent actionable content, particularly with its use of visual aids, checklists, and lists of questions to ask a physician. These seemingly simple interventions are very effective in changing how AI chatbots present information and can be impactful for patient care.

Importantly, our work provides validation of prompt engineering and facilitates urologist's leveraging AI chatbots for clinical purposes. AI chatbots can serve a lot of useful roles, such as summarizing a patient's condition, providing discharge instructions, or helping prepare PEMs. The principles of prompt engineering are as follows: to be specific, provide context around a question, experiment with different prompts/iterate and refine those prompts, identify specific goals, asking it to play a role, asking open-ended questions, requesting examples, and giving a specific temporal range.[24] In developing healthcare specific AI chatbots, these principles should be combined with our findings to present PEMs as effectively as possible to patients. With ever increasing demands on the urologist with upwards of 80% experiencing burnout, effectively using various tools at our disposal to help address these increased workloads is critical.[25]

Our study is not without its limitations. For instance, the questions evaluated were general frequently asked questions derived from publicly available resources, including top urology program websites and American Urological Association patient education materials, which may reflect clinician-driven priorities rather than the specific language, concerns, or informational needs of patients. As such, these questions may not fully capture how patients would naturally interact with a generative AI model in real-world settings, including the use of personalized details, iterative clarification, conversational phrasing, or informal prompt refinement when responses are difficult to understand. Relatedly, this study did not assess the emotional impact of AI-generated responses, such as whether answers might provide reassurance or contribute to patient anxiety. Moreover, as ChatGPT is an ever-changing model that is constantly learning and updating its responses, our study is specific to the period that questions were posed to it. To reduce memory bias in this study, all questions were posed at one time and collected in an offline document so that any temporal change in answers would not impact this study. In development and validation of prompt engineering specific to PEMs, we are still currently limited to reconstructive urology and urologic oncology and should take caution before applying to other fields of Urology or medicine. Nevertheless, we present a variety of questions posed to ChatGPT in three different urologic cancers with excellent responses, showing potential for other cancers and disciplines. Future research should evaluate real patient interactions with an AI chatbot in a conversation across many disciplines in medicine.

CONCLUSIONS

ChatGPT4o responds to patient frequently asked questions in urologic oncology with highly accurate and quality responses at a reading level much higher than what is recommended for patient educational materials. When using prompt engineering, responses remained highly

© 2026 Canadian Urological Association

accurate, while changing to a 5th-7th grade level, which is the recommended reading level of patient educational materials. We show the potential of AI chatbots in providing high quality, accurate, easily accessible patient educational materials in Urologic Oncology.

DRAFT

REFERENCES

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

© 2026 Canadian Urological Association

2. Driessens H, van Wijk L, Buis CI, et al. Low health literacy is associated with worse postoperative outcomes following hepato-pancreato-biliary cancer surgery. *HPB* 2022;24:1869-77. https://doi.org/10.1016/j.hpb.2022.07.006
3. Koay K, Schofield P, Jefford M. Importance of health literacy in oncology. *Asia Pac J Clin Oncol* 2012;8:14-23. https://doi.org/10.1111/j.1743-7563.2012.01522.x
4. Luckenbaugh AN, Moses KA. The impact of health literacy on urologic oncology care. *Urol Oncol* 2022;40:117-9. https://doi.org/10.1016/j.urolonc.2019.06.016
5. Stiller C, Brandt L, Adams M, et al. Improving the readability of patient education materials in physical therapy. *Cureus* 2024;16:e54525.
6. Rooney MK, Santiago G, Perni S, et al. Readability of patient education materials from high-impact medical journals: A 20-year analysis. *J Patient Exp* 2021;8:2374373521998847. https://doi.org/10.1177/2374373521998847
7. Colaco M, Svider PF, Agarwal N, et al. Readability assessment of online urology patient education materials. *J Urol* 2013;189:1048-52. https://doi.org/10.1016/j.juro.2012.08.255
8. Rodler S, Maruccia S, Abreu A, et al. Readability assessment of patient education materials on uro-oncological diseases using automated measures. *Eur Urol Focus* [Internet]. 2024 Jul 23 [cited 2024 Aug 27]; Available from: https://www.sciencedirect.com/science/article/pii/S2405456924001172
9. Giakas JA, Zaliznyak M, Kohut-Jackson A, et al. Quality and readability of online Health information on common urologic cancers: Assessing barriers to health literacy in urologic oncology. *Urol Pract* 2024;11:670-6. https://doi.org/10.1097/UPJ.0000000000000574
10. Pan A, Musheyev D, Bockelman D, et al. Assessment of artificial intelligence chatbot responses to top searched queries about cancer. *JAMA Oncol* 2023;9:1437-40. https://doi.org/10.1001/jamaoncol.2023.2947
11. Sivanesan N, Diaz GM, Kandala K, et al. Why prompting matters: Achieving clinically accurate and consistent responses with Chat GPT. *BJU Int* 2025. https://doi.org/10.1111/bju.16738
12. Hello GPT-4o [Internet]. [cited 2024 Aug 27]. Available from: https://openai.com/index/hello-gpt-4o/
13. Charnock D, Shepperd S, Needham G, et al. DISCERN: An instrument for judging the quality of written consumer health information on treatment choices. *J Epidemiol Community Health* 1999;53:105-11. https://doi.org/10.1136/jech.53.2.105
14. Shoemaker SJ, Wolf MS, Brach C. Development of the patient education materials assessment Tool (PEMAT): A new measure of understandability and actionability for print and audiovisual patient information. *Patient Educ Couns* 2014;96:395-403. https://doi.org/10.1016/j.pec.2014.05.027
15. Wang LW, Miller MJ, Schmitt MR, et al. Assessing readability formula differences with written health information materials: Application, results, and recommendations. *Res Soc Adm Pharm RSAP* 2013;9:503-16. https://doi.org/10.1016/j.sapharm.2012.05.009
16. Meskó B. Prompt engineering as an important emerging skill for medical professionals: Tutorial. *J Med Internet Res* 2023;25:e50638. https://doi.org/10.2196/50638
17. Alshak MN, Cecelic J, Florissi I, et al. Assessing ChatGPT responses to frequently asked patient questions in reconstructive urology. Urol Pract [Internet]. 2024 Dec 30 [cited 2025 Feb 21]; Available from: https://www.auajournals.org/doi/10.1097/UPJ.0000000000000792

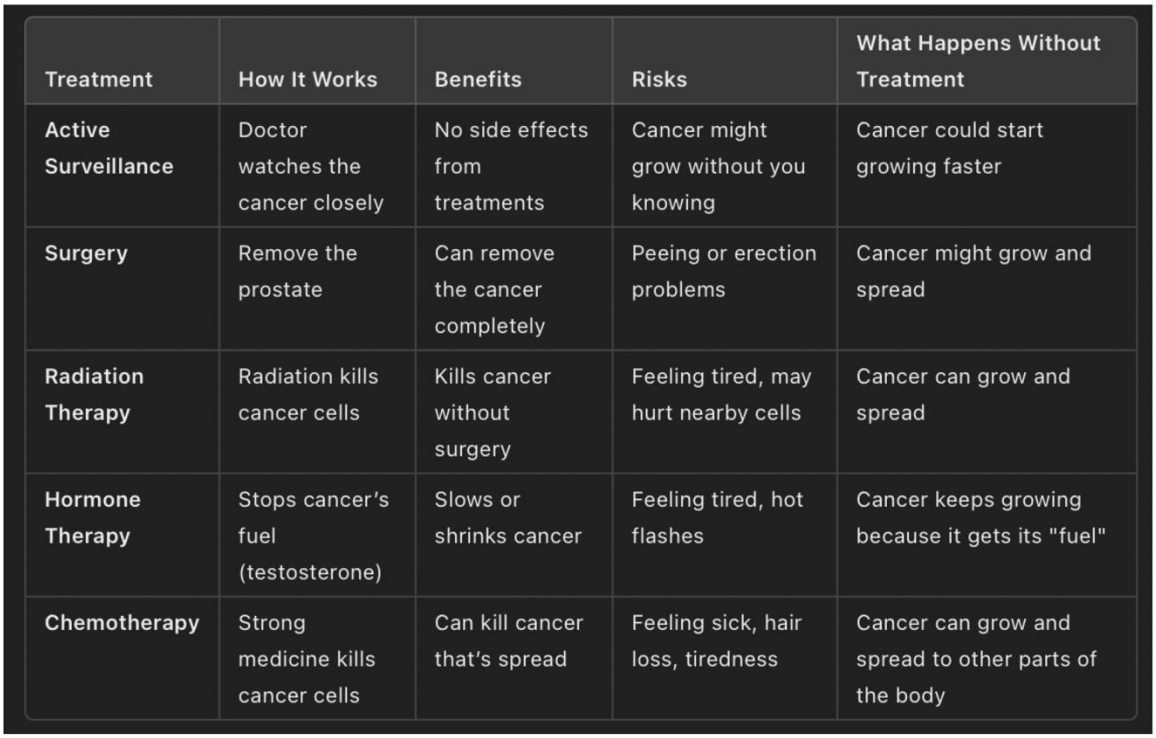
© 2026 Canadian Urological Association

18. de Looper M, van Weert JCM, Schouten BC, et al. The Influence of online health information seeking before a consultation on anxiety, satisfaction, and information recall, mediated by patient participation: Field study. *J Med Internet Res* 2021 Jul 5;23:e23670. https://doi.org/10.2196/23670
19. Bhattad PB, Pacifico L. Empowering patients: Promoting patient education and health literacy. *Cureus* 2022;14:e27336. https://doi.org/10.7759/cureus.27336
20. Nedbal C, Naik N, Castellani D, et al. ChatGPT in urology practice: Revolutionizing efficiency and patient care with generative artificial intelligence. *Curr Opin Urol* 2024;34:98-104. https://doi.org/10.1097/MOU.0000000000001151
21. Ayers JW, Poliak A, Dredze M, et al. Comparing physician and artificial intelligence Chatbot responses to patient questions posted to a public social media forum. *JAMA Intern Med* 2023;183:589-96. https://doi.org/10.1001/jamainternmed.2023.1838
22. Papadakos JK, Giannopoulos E, McBain S, et al. Quality assessment of cancer patient education materials: The current state of systemic therapy patient education in fourteen cancer centres across Ontario, Canada. *Multicenter Study* 2021;29:3513-9. https://doi.org/10.1007/s00520-020-05859-2
23. Mitus AJ, Coughlin L. The value of actionable content in a clinical setting: Access to better information facilitates enhanced cancer care. *Am Health Drug Benefits* 2013;6:104-6.
24. Meskó B. Prompt engineering as an important emerging skill for medical professionals: Tutorial. *J Med Internet Res* 2023;25:e50638. https://doi.org/10.2196/50638
25. Shoureshi P, Guerre M, Seideman CA, et al. Addressing burnout in urology: A qualitative assessment of interventions. *Urol Pract* 2022;9:101-7. https://doi.org/10.1097/UPJ.0000000000000282

© 2026 Canadian Urological Association

FIGURES AND TABLES

Figure 1. An example of a visual created at the end of a response by CHATGPT after prompting engineering in response to "How do you treat prostate cancer?"



Treatment	How It Works	Benefits	Risks	What Happens Without Treatment
Active Surveillance	Doctor watches the cancer closely	No side effects from treatments	Cancer might grow without you knowing	Cancer could start growing faster
Surgery	Remove the prostate	Can remove the cancer completely	Peeing or erection problems	Cancer might grow and spread
Radiation Therapy	Radiation kills cancer cells	Kills cancer without surgery	Feeling tired, may hurt nearby cells	Cancer can grow and spread
Hormone Therapy	Stops cancer's fuel (testosterone)	Slows or shrinks cancer	Feeling tired, hot flashes	Cancer keeps growing because it gets its "fuel"
Chemotherapy	Strong medicine kills cancer cells	Can kill cancer that's spread	Feeling sick, hair loss, tiredness	Cancer can grow and spread to other parts of the body

Table 1. Evaluation of ChatGPT-4o responses to search queries related to prostate, bladder, and kidney cancer

Measure	Overall responses	Prostate cancer	Bladder cancer	Kidney cancer
Number of questions	45	15	15	15
DISCERN score (median, IQR)[a]	4 (1-5)	4 (1-4)	4 (1-5)	4 (1-5)
DISCERN score (mean, standard deviation)[a]	3.42±1.8	3.41±1.80	3.38±1.82	3.47±1.78
PEMAT-P Understandability score % (mean, standard deviation)[b]	88±4	92±6	86 ± 7	87±3

© 2026 Canadian Urological Association

PEMAT-P Actionability score % (mean, standard deviation)[b]	26±20	14±19	24±27	36±29
Misinformation (mean, range)[c]	1.19 (1–3)	1.23 (1–3)	1.17 (1–2)	1.17 (1–2)
Major misinformation (% of questions)[d]	4%	7%	0%	0%
Flesch-Kincaid Grade Level (median, range)[e]	10.5 (6.2–12.7)	10.5 (6.2–12.7)	11.1 (8.1-12.6)	9.8 (7.2–11.9)
Flesch-Kincaid Reading level (median, range)[f]	42 (25.4–69.2) College level	42 (25.4–69.2) College	39.2 (25.7–59) College	49.2 (32.8-–66.6) College
Word Count (median, range)	1058 (254–1431)	947 (254–1196)	1043 (607–1272)	1186 (920–1431)

[a]Measured the quality of health consumer information, scored from 1 (low) to 5 (high). [b]Patient Education Materials Assessment Tool for Printable Materials, Scored from 0% (low) to 100% (high). [c]Scored from 1 (no misinformation) to 5 (high misinformation). [d]Incorrect or missing major diagnostic or treatment options, significant errors, 3 or higher score. [e]Scored from 0 (easy to read) to 18 (most difficult to read), representing U.S. grade levels. [f]Scored from 100-90 (5th grade level) to 10-0 (professional level). [*]Statistically significant

Table 2. Evaluation of ChatGPT-4o responses to search queries related to prostate, bladder, and kidney cancer before and after prompt engineering

Measure	**Original responses**	**Prompt engineering arm**	**p**
Number of questions	45	45	
DISCERN score (median, IQR)[a]	4 (1–5)	5 (5–5)	<0.0001[*]
DISCERN score (mean, standard deviation)[a]	3.42±1.8	4.47±1.1	<0.0001[*]
PEMAT-P Understandability score % (mean, standard deviation)[b]	88±4	96±1	<0.0001[*]

© 2026 Canadian Urological Association

PEMAT-P Actionability score % (mean, standard deviation)[b]	26±20	84±5	<0.0001*
Misinformation (mean, range)[c]	1.19 (1–3)	1.34 (1–3)	0.06
Major misinformation (% of questions)[d]	4%	7%	0.99
Flesch-Kincaid grade level (median, range)[e]	10.5 (6.2–12.7)	5.3 (2.9–7.0)	<0.0001*
Flesch-Kincaid reading level (median, range)[f]	42 (25.4–69.2) College level	71.7 (61.9–88.6) 7th Grade level	<0.0001*
Word count (median, range)	1058 (254–1431)	705 (576–911)	<0.0001*

[a]Measured the quality of health consumer information, scored from 1 (low) to 5 (high). [b]Patient Education Materials Assessment Tool for Printable Materials, Scored from 0% (low) to 100% (high). [c]Scored from 1 (no misinformation) to 5 (high misinformation). [d]Incorrect or missing major diagnostic or treatment options, significant errors, 3 or higher score. [e]Scored from 0 (easy to read) to 18 (most difficult to read), representing U.S. grade levels. [f]Scored from 100-90 (5th grade level) to 10-0 (professional level). *Statistically significant

DRAFT

© 2026 Canadian Urological Association