Am J Clin Exp Urol 2025;13(2):176-185
www.ajceu.us /ISSN:2330-1910/AJCEU0165231

Original Article

DeepSeek vs ChatGPT: a comparison study of their performance in answering prostate cancer radiotherapy questions in multiple languages

Peng-Wei Luo[1,3*], Ji-Wen Liu[2*], Xi Xie[3*], Jia-Wei Jiang[1,4], Xin-Yu Huo[3], Zhen-Lin Chen[1], Zhang-Cheng Huang[1,5], Shao-Qin Jiang[1], Meng-Qiang Li[1]

[1]Department of Urology, Fujian Union Hospital, Fujian Medical University, Fuzhou, Fujian, China; [2]Department of Urology, The General Hospital of Western Theater Command, Chengdu, Sichuan, China; [3]Department of Urology, The First Affiliated Hospital of Chengdu Medical College, Chengdu, Sichuan, China; [4]Department of Urology, Affiliated Jinhua Hospital, Zhejiang University School of Medicine, Jinhua, Zhejiang, China; [5]Department of Urology, The Second People's Hospital Affiliated to Fujian University of Traditional Chinese Medicine, Fuzhou, Fujian, China. [*]Equal contributors.

Received April 10, 2025; Accepted April 25, 2025; Epub April 25, 2025; Published April 30, 2025

Abstract: Introduction: The medical information generated by large language models (LLM) is crucial for improving patient education and clinical decision-making. This study aims to evaluate the performance of two LLMs (DeepSeek and ChatGPT) in answering questions related to prostate cancer radiotherapy in both Chinese and English environments. Through a comparative analysis, we aim to determine which model can provide higher-quality answers in different language environments. Methods: A structured evaluation framework was developed using a set of clinically relevant questions covering three key domains: foundational knowledge, patient education, and treatment and follow-up care. Responses from DeepSeek and ChatGPT were generated in both English and Chinese and independently assessed by a panel of five oncology specialists using a five-point Likert scale. Statistical analyses, including the Wilcoxon signed-rank test, were performed to compare the models' performance across different linguistic contexts. Results: This study ultimately included 33 questions for scoring. In Chinese, DeepSeek outperformed ChatGPT, achieving top ratings (score = 5) in 75.76% vs. 36.36% of responses ($P < 0.001$), excelling in foundational knowledge (76.92% vs. 38.46%, $P = 0.047$) and treatment/follow-up (81.82% vs. 36.36%, $P = 0.031$). In English, ChatGPT showed comparable performance (66.7% vs. 54.55% top-rated responses, $P = 0.236$), with marginal advantages in treatment/follow-up (63.64% vs. 54.55%, $P = 0.563$). DeepSeek maintained strengths in English foundational knowledge (69.23% vs. 30.77%, $P = 0.047$) and patient education (88.89% vs. 55.56%, $P = 0.125$). These findings underscore DeepSeek's superior Chinese proficiency and language-specific optimization impacts. Conclusions: This study shows that DeepSeek performs excellently in providing Chinese medical information, while the two models perform similarly in an English environment. These findings underscore the importance of selecting language-specific artificial intelligence (AI) models to enhance the accuracy and reliability of medical AI applications. While both models show promise in supporting patient education and clinical decision-making, human expert review remains necessary to ensure response accuracy and minimize potential misinformation.

Keywords: Artificial intelligence, DeepSeek, ChatGPT, prostate cancer, radiotherapy

Introduction

Prostate cancer is one of the leading cancers affecting men globally, with high incidence and mortality rates. As the most prevalent malignancy in men, prostate cancer presents a major healthcare challenge, particularly in older populations [1]. Advances in early detection by prostate-specific antigen (PSA) screening and treatment strategies such as surgery, radiotherapy, and hormone therapy, have significantly improved the prognosis for patients [2, 3]. Radiotherapy, in particular, plays a pivotal role in the management of localized and locally advanced prostate cancer, offering outcomes comparable to surgery in certain patient cohorts [4, 5]. While external beam radiation therapy (EBRT) and brachytherapy are widely used,

the precision and effectiveness of these treatments are highly dependent on accurate diagnosis, treatment planning, and patient education [6]. Thus, ensuring that patients and healthcare providers have access to accurate and comprehensible information is critical to improving treatment outcomes.

With the increasing complexity of medical care, Artificial Intelligence (AI), particularly large language models (LLMs), is becoming an essential tool in modern healthcare [7]. These models, powered by advanced natural language processing (NLP) and machine learning algorithms, have shown promise in various medical applications, including diagnostics, treatment planning, and patient education [8]. AI-driven systems such as ChatGPT have been leveraged to provide medical information and support clinical decision-making. However, while LLMs like ChatGPT offer versatile applications, their performance may vary based on language, domain expertise, and specific medical contexts [7, 9, 10]. This variation raises the question of how different models perform when tasked with providing detailed medical knowledge in oncology.

A relatively recent and powerful generative AI model, DeepSeek, has emerged as an alternative to traditional LLMs [11, 12]. DeepSeek is designed with advanced reasoning capabilities that allow it to engage in complex problem-solving tasks with greater transparency and clarity. Its open-source nature and adaptive approach, powered by a Mixture of Experts framework, enable it to perform well across a wide range of medical domains, including oncology [13]. This model offers potential advantages in clinical settings, where both accuracy and interpretability of responses are critical. DeepSeek's focus on transparency allows healthcare providers to better understand the reasoning behind AI-generated recommendations, potentially increasing clinician trust and enhancing patient education [14].

In the context of prostate cancer radiotherapy, the need for effective patient education is crucial, as patients must understand the nature of their disease, treatment options, potential side effects, and the long-term management plan [6]. As such, this study compares the performance of DeepSeek and ChatGPT in answering questions related to prostate cancer radiotherapy. Specifically, we aim to assess how these AI models handle foundational knowledge, patient education, and treatment-related questions in both English and Chinese contexts. By evaluating the accuracy, comprehensiveness, and clarity of their responses, this study seeks to determine which AI model is more suitable for supporting healthcare professionals and patients in making informed decisions regarding prostate cancer treatment.

Materials and methods

Study design

This study was designed to evaluate the performance of two LLMs, i.e., DeepSeek R1 and ChatGPT-4o, in providing information on prostate cancer radiotherapy. A comprehensive set of evaluation questions was created using established clinical treatment guidelines and common patient queries observed in clinical settings. During the question refinement process, the research team carefully examined and removed questions that were redundant, unclear, or overly subjective, ensuring both clarity and clinical applicability. The final set of questions was organized into three main areas: foundational knowledge, patient education, and treatment and follow-up care. This structure enabled a thorough evaluation of the models' performance from various angles, while maintaining relevance to real-world clinical settings.

Assessment process

Evaluation of DeepSeek R1 and ChatGPT-4o was conducted using their respective versions dated in March 2025, with responses generated in both English and Chinese. Each question was entered independently using the "New Chat" function to ensure that prior interactions did not influence the model's responses. Example prompts included "We are conducting a clinical consultation for a prostate cancer patient. Please answer the following questions about prostate cancer". This standardized approach aimed to assess the accuracy and comprehensiveness of each model's responses in the context of prostate cancer radiotherapy. Since this study does not involve patients

Table 1. Definition of a five-point Likert scale for scoring responses to DeepSeek and ChatGPT

	Definition
5	The information is comprehensive and accurate, covering all relevant aspects, and is highly reliable, which can be fully trusted.
4	The information is relatively accurate and comprehensive, covering most of the key aspects, and can meet most of the needs.
3	The accuracy and comprehensiveness of the information basically meet the requirements, but still need to be improved.
2	There are some problems with the accuracy and completeness of the information, which need to be further supplemented or modified.
1	The information contains a lot of errors and omissions.

or their private data, approval from an ethical committee was not necessary.

To guarantee a thorough and clinically pertinent evaluation, a panel of five oncology experts with extensive experience in prostate cancer radiotherapy was formed. The panel consisted of one junior specialist, two mid-level specialists, and two senior specialists. Each group member independently evaluated the responses generated by the model using a five-point Likert scale, assessing the accuracy, relevance, and completeness of the information provided. All group members were kept blind to which LLMs provided the answers. Any scoring discrepancies were addressed through facilitated discussions among the panel members. Each answer was finally evaluated by a senior clinician who has more than a decade of experience in prostate cancer treatment, ensuring consistency and methodological rigor throughout the evaluation process (**Table 1**).

Statistical analysis

Descriptive statistics were used to summarize the evaluation results, with frequency and percentage distributions calculated for categorical data. The Wilcoxon signed-rank test was employed to compare the accuracy and comprehensiveness of responses generated by DeepSeek R1 and ChatGPT-4o. All statistical analyses and graphical representations were performed using GraphPad Prism (version 8.0.2). The significance of difference between experimental groups was determined using the two-tailed Student's t-test. Mann-Whitney U-test was used for data that did not conform to a normal distribution. $P < 0.05$ indicated significance.

Results

The same set of questions yields different response qualities from LLMs due to different languages

After screening the initially questions, a total of 33 questions were finally included in the questions & answers (Q&A) scoring (**Figure 1**). The questions in the study were evaluated using a standardized Likert scale (**Table 1**). The results revealed significant differences in the performance of DeepSeek and ChatGPT across across various language contexts. Although both models scored the same on nearly half of the questions (15/33) in a Chinese context, DeepSeek performed better than ChatGPT on 16 of the 33 questions (**Figure 2A**). In contrast, the performance of DeepSeek and ChatGPT was more comparable in the English-language setting. Among the 33 evaluated responses, both models received identical scores on 13 questions, while the score difference was within one point for 15 responses (**Figure 2B**). These results suggest that the two models demonstrate relatively comparable proficiency in processing English-language medical queries.

Comparison of response qualities between DeepSeek and ChatGPT on prostate cancer radiotherapy in Chinese and English contexts

Responses from DeepSeek and ChatGPT on prostate cancer radiotherapy were evaluated in both Chinese and English contexts using a five-point Likert scale. The analysis revealed significant differences in performance between the two models across different languages. In the Chinese context, DeepSeek outperformed

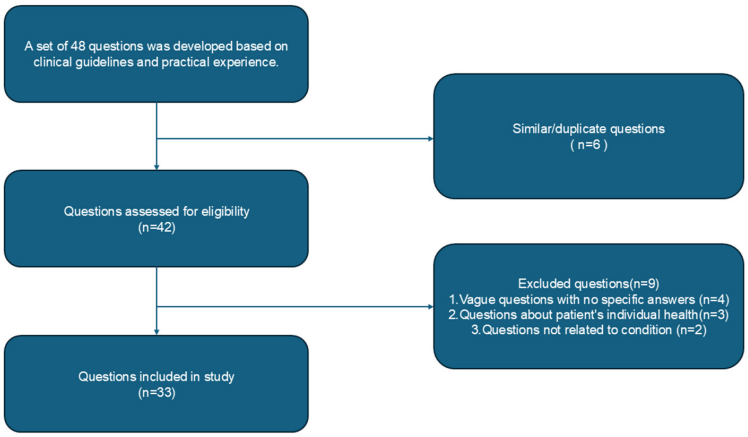

Figure 1. Selection flowchart of prostate cancer radiotherapy questions. Frequently asked questions about the knowledge and management of prostate cancer radiotherapy are derived from existing clinical treatment guidelines and patients' questions in actual treatment.

ChatGPT, with 75.76% of responses receiving the highest rating (score = 5) compared to 36.36% for ChatGPT (**Table 2**, $P < 0.001$). Conversely, ChatGPT performed slightly better in the English context, with 66.7% of responses rated as 5 compared to 54.55% for DeepSeek (**Table 2**, $P = 0.236$). These results show that DeepSeek performs excellently in the Chinese medical context, while in the English environment, there is no significant difference in the answer qualities of the two models. This may reflect that language training and optimization may bring differences to the model performance (**Figure 3**).

Comparison of DeepSeek and ChatGPT performance in prostate cancer radiotherapy related questions across Chinese contexts

An evaluation of the responses from DeepSeek and ChatGPT on foundational knowledge, patient education, and treatment and follow-up in prostate cancer radiotherapy showed significant performance differences (**Figure 4**). For fundamental knowledge, DeepSeek demonstrated superior performance with 76.92% of its responses receiving the highest rating, compared to 38.46% for ChatGPT (**Table 3**, $P = 0.047$). ChatGPT exhibited a more evenly distributed rating profile with a higher proportion of responses rated 4 (30.77%) and 3 (23.08%) compared to DeepSeek (15.38% and 7.69%).

In the patient education category, DeepSeek performed better than ChatGPT in the 5-point rating, with 66.67% of the responses being rated as 5 points, while ChatGPT received this rating in 33.33% of the responses. Although DeepSeek received a higher proportion of high-scoring responses, ChatGPT received a larger proportion of 4-point responses than DeepSeek (44.44% vs. 22.22%). Overall, there was no statistically significant difference in their overall performance (**Table 3**, $P = 0.375$).

For treatment and follow-up, DeepSeek's performance was significantly higher, with 81.82% of responses receiving a rating of 5, compared to only 36.36% for ChatGPT (**Table 3**, $P = 0.031$). Notably, ChatGPT exhibited a more balanced rating distribution, with 36.36% of responses rated 4, while no responses from DeepSeek received this rating. Additionally, a small proportion of responses from both models were rated 3, while only ChatGPT received a 2 rating (9.09%).

Comparison of DeepSeek and ChatGPT performance in prostate cancer radiotherapy related questions across English contexts

Our team further consulted DeepSeek and ChatGPT in English context to evaluate their performance in prostate cancer radiation therapy-related knowledge. The results showed that the performance gap between the two had narrowed. In the domain of foundational knowledge (n = 13), DeepSeek exhibited a higher proportion of top-rated responses, with 69.23% of its responses receiving a rating of 5, compared to 30.77% for ChatGPT (**Table 4**, $P = 0.047$). Additionally, 38.46% of DeepSeek's responses were rated 4, whereas ChatGPT received only 15.38% in this category. These findings indicate that DeepSeek provided more comprehensive and highly rated responses in this category.

For patient education (n = 9), DeepSeek's responses were rated 5 in 88.89% of responses, whereas ChatGPT achieved this rating in 55.56% of responses (**Table 4**, $P = 0.125$). While DeepSeek demonstrated a higher pro-

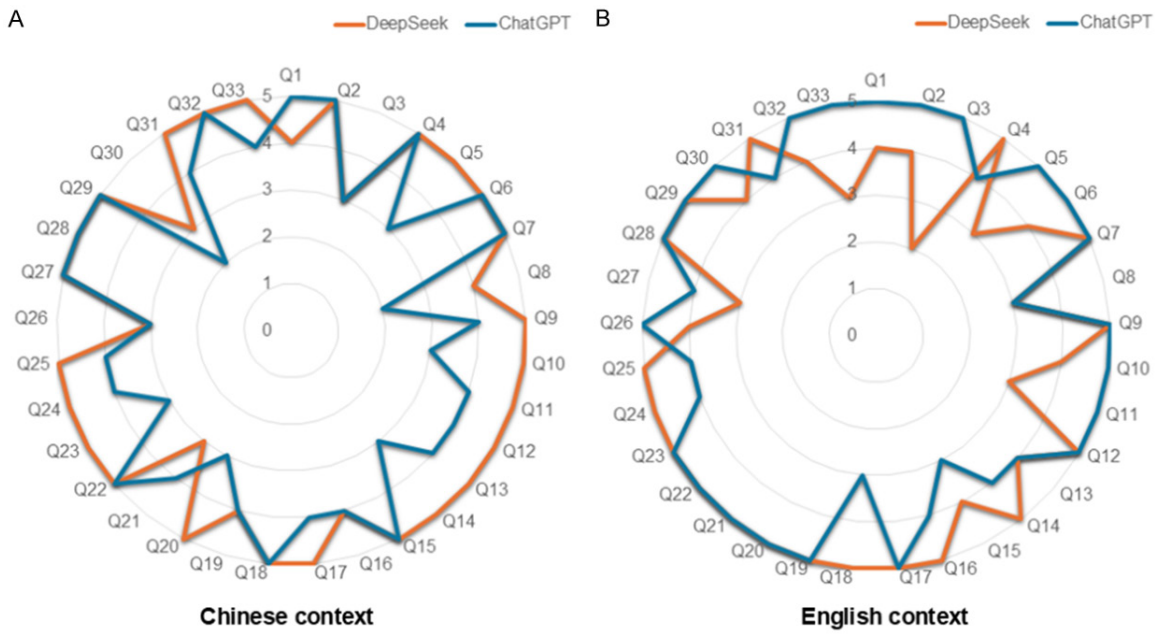

Figure 2. A radar chart showing the performance of DeepSeek (orange) and ChatGPT (blue) in both Chinese and English contexts.

Table 2. Evaluation of DeepSeek and ChatGPT's responses to questions related to prostate cancer radiotherapy in different language contexts

	Five-point Likert scale	DeepSeek	ChatGPT	*P-value*
Chinese context (n = 33)	5	25 (75.76%)	12 (36.36%)	<0.001
	4	4 (12.12%)	12 (36.36%)	
	3	4 (12.12%)	7 (21.21%)	
	2	0 (0.0%)	2 (6.06%)	
	1	0 (0.0%)	0 (0.0%)	
English context (n = 33)	5	18 (54.55%)	22 (66.70%)	0.236
	4	9 (27.27%)	8 (24.24%)	
	3	5 (15.15%)	3 (9.09%)	
	2	1 (3.03%)	0 (0.0%)	
	1	0 (0.0%)	0 (0.0%)	

portion of top-rated responses, the difference was not statistically significant. In the treatment and follow-up (n = 11) category, ChatGPT received a slightly higher proportion of 5 ratings (63.64%) compared to DeepSeek (54.55%) (**Table 4**, $P = 0.563$). However, the difference was minimal and not statistically significant.

Overall, these results suggest that DeepSeek demonstrated superior performance in foundational knowledge and patient education, whereas ChatGPT provided slightly better responses in the treatment and follow-up category. However, the differences observed in patient education and treatment-related responses did not reach statistical significance, indicating that the performance of these two models in these areas was generally quite similar in English context (**Figure 5**).

Discussion

With the development of LLMs, their applications in the healthcare sector have attracted widespread attention [7]. Among them, ChatGPT stands out as one of the most representative models. Previous studies have shown that, in diagnosing inflammatory rheumatic diseas-

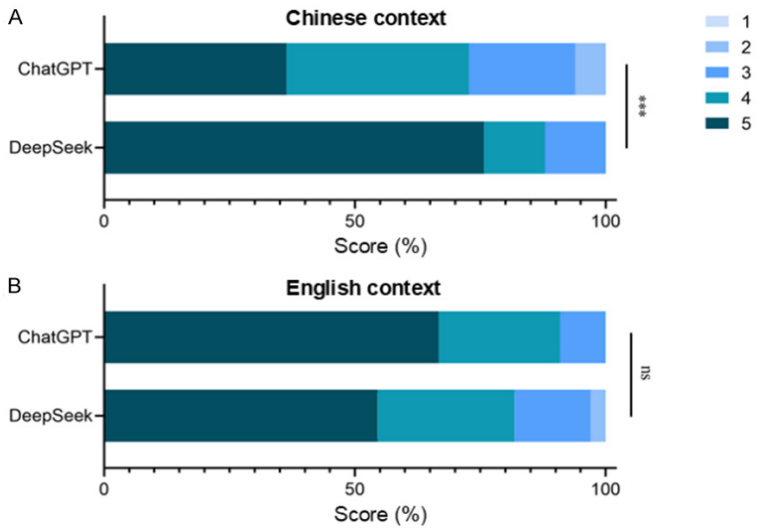

Figure 3. A bar chart illustrating the ratings of responses provided by DeepSeek and ChatGPT, evaluated on a five-point Likert scale in both Chinese and English contexts. The horizontal axis scale (1-100) represents the percentage distribution of scores across different levels on the five-point Likert scale. ns, not significant; ***$P < 0.001$.

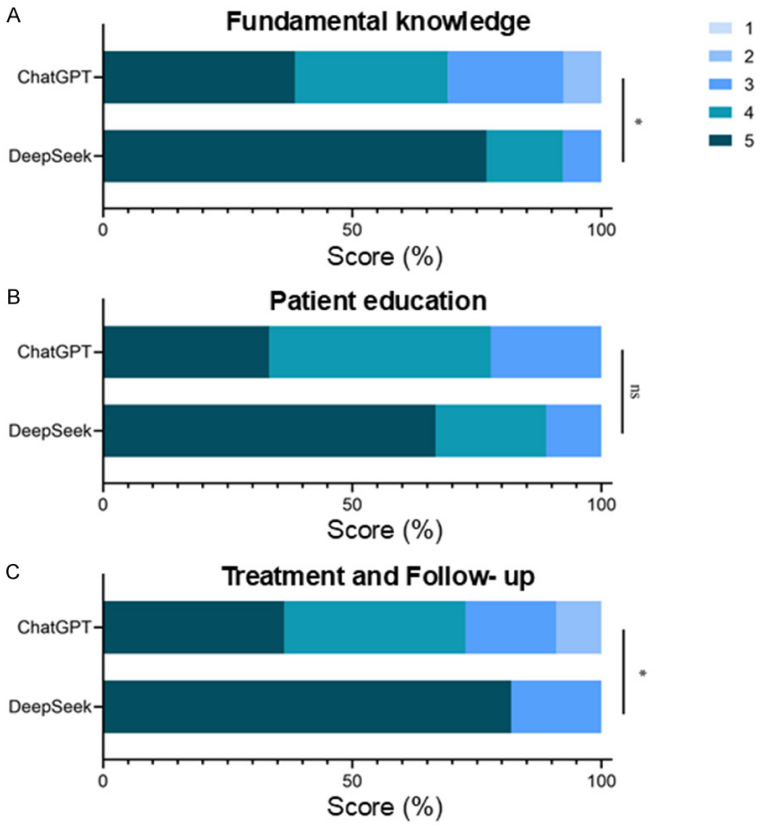

Figure 4. A bar chart illustrating the ratings of responses related to foundational knowledge, patient education, and treatment and follow-up questions, evaluated using a five-point Likert scale in the Chinese context. The horizontal axis scale (1-100) represents the percentage distribution of scores across different levels on the five-point Likert scale. ns, not significant; *$P < 0.05$.

es, ChatGPT demonstrated diagnostic accuracy and effectiveness comparable to that of professional rheumatologists. Even though ChatGPT may not be as accurate as clinicians in some details, it can still provide general information [15]. LLMs have also shown potential in patient health education. Studies on pulmonary nodules have shown that ChatGPT can appropriately answer most related questions without obvious prompts [16]. Additionally, when used as an educational resource for patients with multiple myeloma, ChatGPT achieved a response accuracy rate of 95%. It indicates that LLMs may be a reliable tool in the field of medical and health education [17].

In a recent development, DeepSeek R1, an LLM developed by a Chinese research team, has garnered global attention due to its outstanding performance and low training costs [18]. It is the first Chinese LLM to perform at a level comparable to ChatGPT-4o [14]. This study comprehensively analyzed the response performance of DeepSeek and ChatGPT in different language environments. Notably, this research is the first to introduce DeepSeek into the field of prostate cancer radiotherapy and to directly compare it with ChatGPT. The results suggest that the performance differences between DeepSeek and ChatGPT emphasize the challenges of language adaptability in LLMs. In the Chinese context, DeepSeek demonstrated a significant advantage (75.76% vs. 36.36% highest rating), likely due to its open-source Mixture of Experts framework and optimization

Table 3. The scoring of DeepSeek and ChatGPT's responses to questions related to prostate cancer radiotherapy in Chinese

	Five-point Likert scale	DeepSeek	ChatGPT	*P-value*
Fundamental knowledge (n = 13)	5	10 (76.92%)	5 (38.46%)	0.047
	4	2 (15.38%)	4 (30.77%)	
	3	1 (7.69%)	3 (23.08%)	
	2	0 (0.0%)	1 (7.69%)	
	1	0 (0.0%)	0 (0.0%)	
Patient education (n = 9)	5	6 (66.67%)	3 (33.33%)	0.375
	4	2 (22.22%)	4 (44.44%)	
	3	1 (11.11%)	2 (22.22%)	
	2	0 (0.0%)	0 (0.0%)	
	1	0 (0.0%)	0 (0.0%)	
Treatment and follow-up (n = 11)	5	9 (81.82%)	4 (36.36%)	0.031
	4	0 (0.0%)	4 (36.36%)	
	3	2 (18.18%)	2 (18.18%)	
	2	0 (0.0%)	1 (9.09%)	
	1	0 (0.0%)	0 (0.0%)	

Table 4. The scoring of DeepSeek and ChatGPT's responses to questions related to prostate cancer radiotherapy in English

	Five-point Likert scale	DeepSeek	ChatGPT	*P-value*
Fundamental knowledge (n = 13)	5	4 (30.77%)	9 (69.23%)	0.047
	4	5 (38.46%)	2 (15.38%)	
	3	3 (23.08%)	2 (15.38%)	
	2	1 (7.69%)	0 (0.0%)	
	1	0 (0.0%)	0 (0.0%)	
Patient education (n = 9)	5	8 (88.89%)	5 (55.56%)	0.125
	4	1 (11.11%)	2 (22.22%)	
	3	0 (0.0%)	2 (22.22%)	
	2	0 (0.0%)	0 (0.0%)	
	1	0 (0.0%)	0 (0.0%)	
Treatment and follow-up (n = 11)	5	6 (54.55%)	7 (63.64%)	0.563
	4	3 (27.27%)	4 (36.36%)	
	3	2 (18.18%)	0 (0.0%)	
	2	0 (0.0%)	0 (0.0%)	
	1	0 (0.0%)	0 (0.0%)	

through training on Chinese medical corpora. In contrast, ChatGPT's competitive edge in the English context (66.7% vs. 54.55%) reflects its pre-training on a vast amount of English data. This discrepancy suggests that model performance is not only dependent on algorithm architecture but also heavily influenced by language-specific resources, such as terminology databases and localized clinical guidelines [19].

The way in which models process text also influences the output results. One of the notable features of DeepSeek is its transparent reasoning mechanism, which may enhance the structural and logical coherence of its answers, thereby better mitigating the impact of prompt bias on the responses [20]. Its high ratings in Chinese patient education questions (66.67% receiving the highest score) indicate that it is particularly well-suited to handle content that

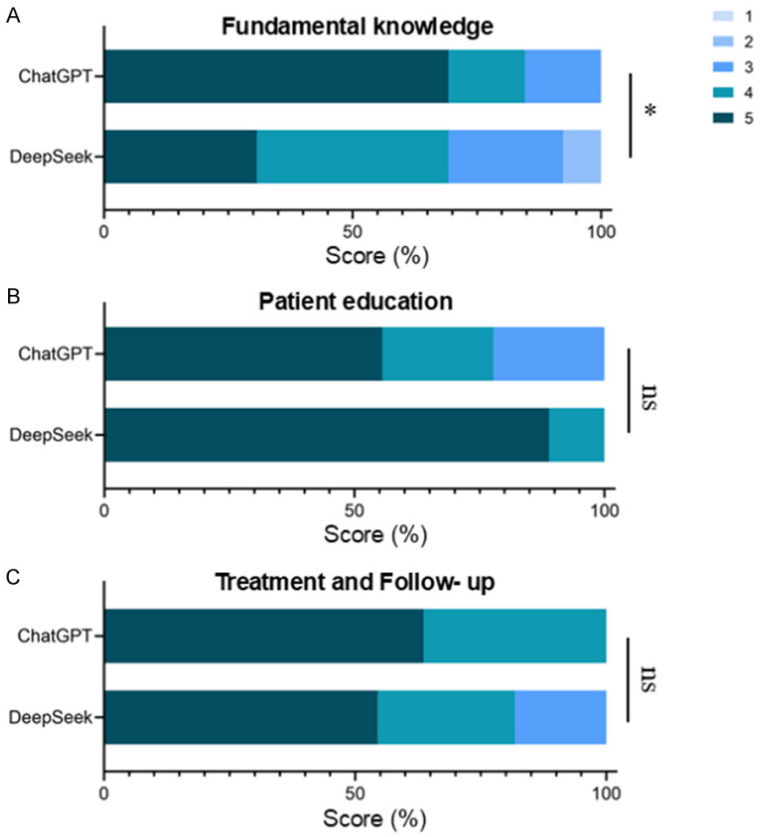

Figure 5. A bar chart illustrating the ratings of responses related to foundational knowledge, patient education, and treatment and follow-up questions, evaluated using a five-point Likert scale in the Englis context. The horizontal axis scale (1-100) represents the percentage distribution of scores across different levels on the five-point Likert scale. ns, not significant; *$P < 0.05$.

requires cultural adaptation, such as the localization of medical terminology. However, ChatGPT also exhibited superior performance in certain areas. For example, in some English-language questions, ChatGPT demonstrated stronger clinical relevance, likely due to its extensive general-purpose corpus and more sophisticated contextual understanding. Nonetheless, its more dispersed ratings (e.g., 44.44% rated 4 in Chinese patient education questions) suggest potential issues with information redundancy or a lack of emphasis on key points.

Linguistic differences directly impact the practical utility of language models [19, 21]. The complexity of Chinese medical queries, particularly regarding term abbreviations and the integration of traditional Chinese and Western medicine concepts, presents additional challenges for model comprehension [10]. However, DeepSeek has been able to overcome some of these difficulties through targeted training on Chinese medical corpora. In contrast, the standardized presentation of medical concepts in English, such as the TNM staging system, provides ChatGPT with a clearer semantic interpretation pathway [22]. Furthermore, aspects such as patients' preferences for treatment modalities, including knowledge of traditional Chinese medicine, may not be fully captured by general-purpose models, necessitating further fine-tuning with localized data for optimization [23, 24]. Consequently, selecting the appropriate large language model is crucial for patients to receive the most precise and linguistically adapted medical information, underscoring the importance of language-specific model optimization in AI-driven healthcare applications.

This study demonstrates that both DeepSeek and ChatGPT perform well in answering questions related to prostate cancer radiotherapy. However, limitations still exist, necessitating human expert review to ensure accuracy and reliability. One of the most critical constraints affecting response quality is the issue of "hallucinations" in large language models, requiring vigilance against potential "overconfidence" risks [24]. AI models inherently tend to generate an answer regardless of the certainty of the information, yet they cannot guarantee the evidence level of their sources [25]. To address this, the expert evaluation mechanism employed in this study - comprising independent assessments by five experts followed by a final review by a senior physician - validates the effectiveness of a human-AI collaborative model in enhancing information reliability. Moving forward, future research could explore real-time interactive systems that dynamically integrate AI-generated content with clinical guidelines, allowing for final verification and

annotation by medical professionals. Such advancements may introduce new diagnostic and treatment paradigms in clinical applications, further improving the accuracy and trustworthiness of AI-assisted medical consultations.

Limitations and future directions

This study has certain limitations. The sample size was relatively small (33 questions) and the scope was restricted to prostate cancer radiotherapy. Future research should expand the evaluation to a broader range of diseases, AI models, and linguistic contexts to ensure more comprehensive insights. Additionally, improvements in real-time model updates and internet search capabilities could further enhance response accuracy and reliability. However, ethical considerations, such as data privacy, security, and accountability, remain critical challenges that require further exploration. Future studies should address these issues to refine the role of AI in medical consultations and improve its integration into clinical practice.

Conclusion

This study demonstrates that both DeepSeek and ChatGPT exhibit strong performance in addressing questions related to prostate cancer radiotherapy. DeepSeek performs better in Chinese context than ChatGPT. This advantage is likely attributed to DeepSeek's Mixture of Experts framework and its optimized training on Chinese medical corpora. In contrast, in the English-language setting, both models performed comparably. These findings suggest that DeepSeek may offer a superior experience for Chinese-speaking users, ensuring timely and accurate medical information in that linguistic context.

Acknowledgements

This work was supported by the Foundation of Health Commission of Chengdu (2022071), the Foundation of the Fujian Natural Sciences Foundation (2022J01260), the Foundation of the National Science Foundation of China (82203309), and the Foundation of Jinhua Science and Technology Program (2021-4-005).

Disclosure of conflict of interest

None.

Address correspondence to: Meng-Qiang Li and Shao-Qin Jiang, Department of Urology, Fujian Union Hospital, Fujian Medical University, Fuzhou 350001, Fujian, China. E-mail: limengqiang1125@163.com (MQL); jiang81555767@126.com (SQJ)

References

[1] Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A and Bray F. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2021; 71: 209-249.

[2] Ilic D, Djulbegovic M, Jung JH, Hwang EC, Zhou Q, Cleves A, Agoritsas T and Dahm P. Prostate cancer screening with prostate-specific antigen (PSA) test: a systematic review and meta-analysis. BMJ 2018; 362: k3519.

[3] Siegel RL, Miller KD, Fuchs HE and Jemal A. Cancer statistics, 2022. CA Cancer J Clin 2022; 72: 7-33.

[4] Wallis CJD, Saskin R, Choo R, Herschorn S, Kodama RT, Satkunasivam R, Shah PS, Danjoux C and Nam RK. Surgery versus radiotherapy for clinically-localized prostate cancer: a systematic review and meta-analysis. Eur Urol 2016; 70: 21-30.

[5] Henry A, Pieters BR, André Siebert F and Hoskin P; UROGEC group of GEC ESTRO with endorsement by the European Association of Urology. GEC-ESTRO ACROP prostate brachytherapy guidelines. Radiother Oncol 2022; 167: 244-251.

[6] Hoffman KE, Penson DF, Zhao Z, Huang LC, Conwill R, Laviana AA, Joyce DD, Luckenbaugh AN, Goodman M, Hamilton AS, Wu XC, Paddock LE, Stroup A, Cooperberg MR, Hashibe M, O'Neil BB, Kaplan SH, Greenfield S, Koyama T and Barocas DA. Patient-reported outcomes through 5 years for active surveillance, surgery, brachytherapy, or external beam radiation with or without androgen deprivation therapy for localized prostate cancer. JAMA 2020; 323: 149-163.

[7] Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF and Ting DSW. Large language models in medicine. Nat Med 2023; 29: 1930-1940.

[8] Van Veen D, Van Uden C, Blankemeier L, Delbrouck JB, Aali A, Bluethgen C, Pareek A, Polacin M, Reis EP, Seehofnerová A, Rohatgi N, Hosamani P, Collins W, Ahuja N, Langlotz CP, Hom J,

Gatidis S, Pauly J and Chaudhari AS. Adapted large language models can outperform medical experts in clinical text summarization. Nat Med 2024; 30: 1134-1142.
[9] Cascella M, Montomoli J, Bellini V and Bignami E. Evaluating the feasibility of ChatGPT in healthcare: an analysis of multiple clinical and research scenarios. J Med Syst 2023; 47: 33.
[10] Yuan XT, Shao CY, Zhang ZZ and Qian D. Comparing the performance of ChatGPT and ERNIE Bot in answering questions regarding liver cancer interventional radiology in Chinese and English contexts: a comparative study. Digit Health 2025; 11: 20552076251315511.
[11] Conroy G and Mallapaty S. How China created AI model DeepSeek and shocked the world. Nature 2025; 638: 300-301.
[12] Normile D. Chinese firm's large language model makes a splash. Science 2025; 387: 238.
[13] Peng Y, Malin BA, Rousseau JF, Wang Y, Xu Z, Xu X, Weng C and Bian J. From GPT to DeepSeek: significant gaps remain in realizing AI in healthcare. J Biomed Inform 2025; 163: 104791.
[14] Kayaalp ME, Prill R, Sezgin EA, Cong T, Królikowska A and Hirschmann MT. DeepSeek versus ChatGPT: multimodal artificial intelligence revolutionizing scientific discovery. From language editing to autonomous content generation-Redefining innovation in research and practice. Knee Surg Sports Traumatol Arthrosc 2025; 33: 1553-1556.
[15] Krusche M, Callhoff J, Knitza J and Ruffer N. Diagnostic accuracy of a large language model in rheumatology: comparison of physician and ChatGPT-4. Rheumatol Int 2024; 44: 303-306.
[16] Mao Y, Xu N, Wu Y, Wang L, Wang H, He Q, Zhao T, Ma S, Zhou M, Jin H, Pei D, Zhang L and Song J. Assessments of lung nodules by an artificial intelligence chatbot using longitudinal CT images. Cell Rep Med 2025; 6: 101988.
[17] Saba L, Fu CL, Khouri J, Faiman B, Anwer F and Chaulagain CP. Evaluating ChatGPT as an educational resource for patients with multiple myeloma: a preliminary investigation. Am J Hematol 2024; 99: 1205-1207.
[18] Gibney E. China's cheap, open AI model DeepSeek thrills scientists. Nature 2025; 638: 13-14.
[19] Shao CY, Li H, Liu XL, Li C, Yang LQ, Zhang YJ, Luo J and Zhao J. Appropriateness and comprehensiveness of using ChatGPT for perioperative patient education in thoracic surgery in different language contexts: survey study. Interact J Med Res 2023; 12: e46900.
[20] Reflections on DeepSeek's breakthrough. Natl Sci Rev 2025; 12: nwaf044.
[21] Li KC, Bu ZJ, Shahjalal M, He BX, Zhuang ZF, Li C, Liu JP, Wang B and Liu ZL. Performance of ChatGPT on Chinese master's degree entrance examination in clinical medicine. PLoS One 2024; 19: e0301702.
[22] Rahsepar AA, Tavakoli N, Kim GHJ, Hassani C, Abtin F and Bedayat A. How AI responds to common lung cancer questions: ChatGPT vs Google Bard. Radiology 2023; 307: e230922.
[23] Guo Y, Wang H, Ren X, Wang T, Chen W, Xu Z and Ge H. Can GPTs accelerate the development of intelligent diagnosis and treatment in traditional Chinese Medicine? A survey and empirical analysis. J Evid Based Med 2025; 18: e70004.
[24] Alkaissi H and McFarlane SI. Artificial hallucinations in ChatGPT: implications in scientific writing. Cureus 2023; 15: e35179.
[25] Ong JCL, Seng BJJ, Law JZF, Low LL, Kwa ALH, Giacomini KM and Ting DSW. Artificial intelligence, ChatGPT, and other large language models for social determinants of health: current state and future directions. Cell Rep Med 2024; 5: 101356.