Research Article

For reprint orders, please contact: info@becaris.com



Journal of **Comparative Effectiveness Research**

Evaluation of large language models for clinical decision support in prostate cancer: a multidisciplinary Delphi study

Ángel Borque-Fernando*[,1], María José Ledo Cepero[2], Antonio Rodríguez-Fernández[3], Estefanía Linares Espinós[4], Sara Martínez Breijo[5], Teresa Alonso Gordoa[6], Fernando López Campos[7], María Espinosa Bosch[8], Eugenia García Fernández[9], Juan Gómez Rivas[10], Joan Carles Vilanova Busquets[11] & Daniel Perez-Fentes[12]

[1]Urology Department, Hospital Universitario Miguel Servet, Universidad de Zaragoza, IIS-Aragón, Zaragoza, Spain
[2]Urology Department, Hospital Universitario Puerta del Mar, Cádiz, Spain
[3]Nuclear Medicine Department, Hospital Universitario Virgen de las Nieves, Granada, Spain; Instituto de Investigación Biosanitaria de Granada, Granada, Spain
[4]Urology Department, Hospital Universitario La Paz, Madrid, Spain
[5]Urology Department, Hospital Universitario de A Coruña, A Coruña, Spain; Epithelial Plasticity & Metastasis Group, Instituto de Investigación Biomédica de A Coruña (INIBIC), A Coruña, Spain
[6]Medical Oncology Department, Hospital Universitario Ramón y Cajal, IRYCIS, Madrid, Spain
[7]Radiotherapy Oncology Department, Hospital Universitario Ramón y Cajal, Madrid, Spain
[8]Pharmacy Department, Hospital Regional Universitario de Málaga, Málaga, Spain
[9]Pathology Department, Hospital Universitario La Paz, Madrid, Spain
[10]Urology Department, Hospital Clínico San Carlos, Madrid, Spain
[11]Radiology Department, Clínica Girona, Girona, Spain
[12]Urology Department, Complejo Hospitalario Universitario de Santiago de Compostela, Santiago de Compostela, Spain
*Author for correspondence: dr.borque@gmail.com

Aim: Large language models (LLMs) may help interpret clinical guideline content. However, their performance in guideline-based prostate cancer (PC) scenarios remains insufficiently validated. This study evaluated the agreement between a multidisciplinary expert panel and responses generated by three leading LLMs (Claude, Gemini and ChatGPT) when addressing structured, guideline-derived clinical questions on PC. **Materials & methods:** A multidisciplinary tumor board created 25 structured questions based on the 2025 European Association of Urology guidelines. The guideline figures and tables were transformed into plain-language summaries for use by the LLMs. Each model was prompted with a simple, unmodified query and a panel of experts in PC management independently rated their level of agreement with the responses. The evaluation used a two-round modified Delphi method with predefined consensus criteria. **Results:** Eleven experts evaluated 25 guideline-based questions. After two Delphi rounds, consensus in agreement was reached for 23/25 responses (92%) with Gemini, 20/25 (80%) with ChatGPT and 14/25 (56%) with Claude ($p = 0.0033$). Explicit disagreement was noted in three questions (12%) with Claude after the second round. Median scores of at least 7 out of 9, indicating strong agreement, were achieved in 24/25 responses (96%) for Gemini, 22/25 (88%) for ChatGPT, and 14/25 (56%) for Claude ($p = 0.0002$). Most consensus difficulties involved complex cases, advanced imaging, genetic testing or questions with limited data or controversial topics. **Conclusion:** Advanced LLMs, when provided with PC guidelines with plain-language adaptations of figures and tables, can achieve strong alignment with expert consensus in routine situations. However, performance varies among models, with Claude showing lower agreement rates, especially in complex cases. These results emphasize the importance of expert oversight, continuous validation and standardized guideline formats interpretable by both clinicians and AI systems.

Plain language summary: Artificial intelligence tools for interpreting prostate cancer guideline questions
What is this article about? Clinical guidelines help doctors make decisions about prostate cancer (PC), but they can be long and complex. Artificial intelligence tools such as Claude, Gemini and ChatGPT may help interpret guideline information. However, the answers generated by these tools need to be checked by medical experts before they can be used in practice.

Becaris PUBLISHING

What methodology was used in this study? PC specialists created 25 questions based on European PC guidelines. The original guideline text was provided to the three artificial intelligence tools, with figures and tables converted into clear, structured text. For each question, experts reviewed the answers and rated whether they agreed with them.
What were the results? Gemini and ChatGPT agreed with expert opinion more often than Claude. After two rounds of expert review, Gemini reached agreement in 23 of 25 answers, ChatGPT in 20 of 25 answers, and Claude in 14 of 25 answers. The tools performed better for routine PC questions than for more complex situations, such as advanced imaging, genetic testing, incomplete clinical information or controversial topics.
Why are these results important? These results suggest that artificial intelligence tools may help interpret information from PC guidelines, especially in straightforward situations. However, their responses are not always reliable and should be checked by healthcare professionals with expertise in PC.

First draft submitted: 14 April 2026; Accepted for publication: 30 September 2026; Published online: 4 November 2026

Keywords: artificial intelligence • clinical decision support systems • consensus • Delphi technique • guideline adherence • large language models • prostate cancer

Prostate cancer (PC) is a major global health concern [1,2]. It is the second most commonly diagnosed cancer in men, with approximately 1.47 million new cases worldwide, corresponding to an age-standardized incidence rate of 29.4 per 100,000 men [1,2]. The disease accounts for nearly 397,000 deaths annually, ranking as the eighth leading cause of cancer mortality globally [1]. In over half of the countries worldwide, PC is the most frequently diagnosed malignancy in men, and in approximately a quarter of all countries, it represents the leading cause of cancer-related death in males [2]. Within Europe, PC is the most commonly diagnosed cancer in men and the third leading cause of cancer death [2]. These data underscore the substantial burden of PC and highlight the need for improved diagnostic and therapeutic strategies.

Clinical practice guidelines, such as those issued by the European Association of Urology (EAU) [2], are essential for standardizing PC management by providing evidence-based recommendations and expert consensus aimed at optimizing patient outcomes. However, their translation into routine clinical practice is often hindered by the complexity of the guidelines, the wide heterogeneity of clinical presentations, frequent updates of recommendations and variability in clinician expertise and institutional resources [3–5]. In this context, large language models (LLMs) have emerged as promising digital tools that may enhance accessibility to current guideline-based recommendations and support real-time clinical decision-making [6,7].

In prior work, we found that submitting simple, unmodified clinical questions to commonly used LLMs yielded responses that aligned more closely with current clinical guidelines than those generated from rephrased or modified prompts [8]. Among the models evaluated, Claude [9] (version Claude-Haiku-2024-22) demonstrated the highest concordance with guideline-based recommendations and expert assessments, outperforming both Gemini [10] (version Gemini-Flash-001, May 2024) and ChatGPT [11] (version GPT-4o-mini-18). Given the rapid pace of development and frequent updates in LLM architectures, continued research is necessary to rigorously evaluate the consistency, accuracy and clinical relevance of LLM-generated recommendations across diverse oncologic scenarios. Ongoing assessment within multidisciplinary frameworks is essential to ensure that these evolving tools are safely and effectively integrated into clinical practice.

This study evaluated the expert-assessed alignment of responses generated by leading LLMs (Claude, Gemini and ChatGPT) when addressing structured, guideline-derived clinical queries related to PC. The primary aim was to determine whether AI-generated responses, produced after providing EAU guidelines with plain-language adaptations of figures and tables, were consistent with expert interpretation of guideline-based recommendations. To achieve this objective, we used a Delphi methodology in which a multidisciplinary panel of specialists independently rated the degree of agreement with LLM-generated responses for key clinical scenarios in PC. This work is part of a larger series examining the application of LLMs across major urological cancers, including renal cell carcinoma, upper tract urothelial carcinoma, muscle-invasive and non-muscle-invasive bladder cancer and penile and testicular cancers.

Materials & methods

Study design

This study, which included a multidisciplinary team of experts in PC, used a two-round Delphi method to reach agreement with responses generated by three LLMs: Claude, Gemini and ChatGPT.

The optimization of LLMs to produce evidence-based responses in accordance with the 2025 EAU guidelines on PC, coupled with their validation through multidisciplinary expert consensus, is part of a broader strategic initiative. This project aims to deliver a registered, AI-driven chatbot (urooncoguias.aeu.es) designed to support clinical decision-making across the full spectrum of uro-oncological conditions.

This project was developed under the coordination of the Spanish Association of Urology (AEU). Its technical implementation was carried out by the Aragón Institute of Health Research (IIS-A) and the Aragón Institute of Engineering Research (I3A). The EAU signed a collaboration agreement with the AEU, granting permission to use its 2025 Guidelines for the development of dedicated LLM models.

Organizational structure & workflow

The project was organized around two separate boards and an expert board panel: the Scientific Board provided strategic oversight and ensured broad representation across all major urological cancer subtypes; The Technical Board provided methodological and digital expertise; The Specialized Tumor Board for PC consisted of multidisciplinary experts, including urologists, radiation oncologists, medical oncologists, nuclear medicine physicians, pathologists, hospital pharmacists and radiologists from relevant Spanish scientific societies: AEU, Spanish Society of Radiation Oncology (SEOR), Genitourinary Alliance for Research and Development (GUARD), Spanish Society of Nuclear Medicine and Molecular Imaging (SEMNIM), Spanish Society of Anatomical Pathology (SEAP), Spanish Society of Hospital Pharmacy (SEFH) and Spanish Society of Medical Radiology (SERAM). This tumor board developed 25 guideline-based clinical questions aligned with the 2025 EAU guidelines [2]. This expert panel of eleven medical specialists independently evaluated the responses generated by LLMs.

Adaptation of clinical guidelines to plain language

All figures and tables from the EAU guidelines for PC [2] were systematically converted into structured, plain-language textual summaries. The guidelines were organized into thematic sections and reviewed by independent members of the specialized PC tumor board. Any differences were settled through consensus among the coordinators, with final approval needing unanimous agreement.

Questionnaire development

The Specialized Tumor Board for PC developed a set of 25 structured questions based on the original EAU guideline content. These addressed key clinical topics such as diagnosis, risk assessment, indications for diagnostic tests and treatment options.

LLM selection & response generation

Three widely accessible LLMs were tested: Claude [12] (Anthropic; version claude-3-5-haiku-20241022, launched 22 October 2024), Gemini [13] (Google; version gemini-2.5-flash-preview-05-20, launched 20 May 2025) and ChatGPT [14] (OpenAI; version gpt-4.1-mini-2025-04-14, launched 14 April 2025). Each model was evaluated using a simple unmodified prompt approach, submitting the 25 structured questions without alteration. The guidelines with plain-language adaptations of figures and tables were provided to enable accurate interpretation by the LLMs.

Delphi validation process & consensus criteria

Expert panelists independently evaluated each LLM response using a 9-point Likert scale (1 = strong disagreement, 9 = strong agreement) to measure alignment with EAU guideline content. Panelists were blinded to the identity of the LLM. They could abstain if they felt unqualified. Ratings from those who participated were included in the analysis.

The Delphi process involved two rounds. In the first round, all responses were scored. Questions lacking consensus were reassessed in a second round, which included anonymized, aggregated feedback summarizing group ratings. Consensus was defined based on three criteria: a median score in the agreement (7–9) or disagreement

(1–3) range; fewer than a third of panelists rating outside these ranges and an interquartile range (IQR) less than 4. Qualitative comments were recorded to contextualize quantitative results.

Statistical analysis

To evaluate differences between models, two binary outcomes were analyzed: consensus in agreement versus no consensus or consensus in disagreement, and median score ≥7 versus <7. Because the same 25 clinical questions were evaluated across all three LLMs, comparisons were treated as paired by question. Overall differences among the three models were assessed using Cochran's Q test. Pairwise comparisons between models were performed using exact McNemar's test. To account for multiple pairwise comparisons, Bonferroni correction was applied within each outcome, resulting in a significance threshold of $p < 0.0167$ for pairwise analyses. Overall tests were considered statistically significant at $p < 0.05$.

The full set of responses generated by each LLM to all 25 questions is available for consultation at https://www.contentednet.com/proyecto-urooncoguias-aeu/

Results

A panel of 11 experts evaluated the responses from the three LLMs (Claude, Gemini, and ChatGPT) to 25 clinical questions on PC management. Since experts could abstain if they felt unqualified to assess a given item, the number of experts who rated each item ranged from 8 to 11.

Median scores assigned by panelists for each question and LLM are shown in Figure 1. The questions, details of the consensus results, number of panelists rating each item, and the summarized panelist comments are shown in Supplementary Table 1. The distribution of scores (1–3 = disagreement, 4–6 = no consensus, 7–9 = agreement) per clinical question and LLM is shown in Supplementary Figure 1.

Consensus in agreement was reached in the first round for 52% (Claude), 80% (Gemini) and 76% (ChatGPT) of the questions. Additional consensus in agreement was achieved in the second round for 4% (Claude), 12% (Gemini) and 4% (ChatGPT). Explicit disagreement was noted in three questions (12%) for Claude after the second round. No consensus was reached for 32% (Claude), 8% (Gemini) and 20% (ChatGPT) (Supplementary Table 1).

After two rounds of assessment, Gemini reached consensus in agreement on 23 of 25 responses (92%), ChatGPT in 20 of 25 responses (80%) and Claude in 14 of 25 responses (56%) (Supplementary Table 1). The analysis of consensus in agreement versus no consensus or consensus in disagreement revealed statistically significant differences among the three models ($p = 0.0033$). Regarding the analysis of median scores ≥7 versus <7, Gemini obtained scores ≥7 in 24 of 25 responses (96%), ChatGPT in 22 of 25 responses (88%) and Claude in 14 of 25 responses (56%) ($p = 0.0002$) (Figure 2).

In the pairwise comparison of models, the number of items reaching consensus in agreement was significantly higher for Gemini compared with Claude ($p = 0.0117$). No significant differences were observed between Claude and ChatGPT ($p = 0.0703$) or Gemini and ChatGPT ($p = 0.2500$) after Bonferroni correction (Figure 2).

When analyzing items with a median score ≥7 versus <7, both Gemini and ChatGPT outperformed Claude ($p = 0.0020$ and $p = 0.0078$, respectively), with no significant difference between Gemini and ChatGPT ($p = 0.5000$) after Bonferroni correction (Figure 2).

Consensus difficulties, requiring additional rounds, disagreement or no agreement, mainly involved managing complex patients, advanced imaging and genetic testing. In four questions, some panelists felt that the question lacked enough data to provide a comprehensive answer (questions 16, 17 and 18) or addressed a controversial topic with no consensus (question 13) (Supplementary Table 1).

Discussion

This study evaluated, under controlled conditions, the performance of three leading LLMs (Claude, Gemini and ChatGPT) in generating responses to structured, guideline-derived clinical questions related to PC. Using a Delphi methodology, a multidisciplinary expert panel systematically assessed their agreement with each model's output. For most questions, responses achieved strong panel support, with Gemini and ChatGPT performing best numerically, followed by Claude. The differences in performance among the models were statistically significant. However, several questions, particularly those involving management of complex clinical cases, advanced imaging, genetic testing or those with insufficient clinical data, failed to achieve strong panel endorsement. In some of these cases, performance could potentially be improved by providing more comprehensive clinical information. These findings

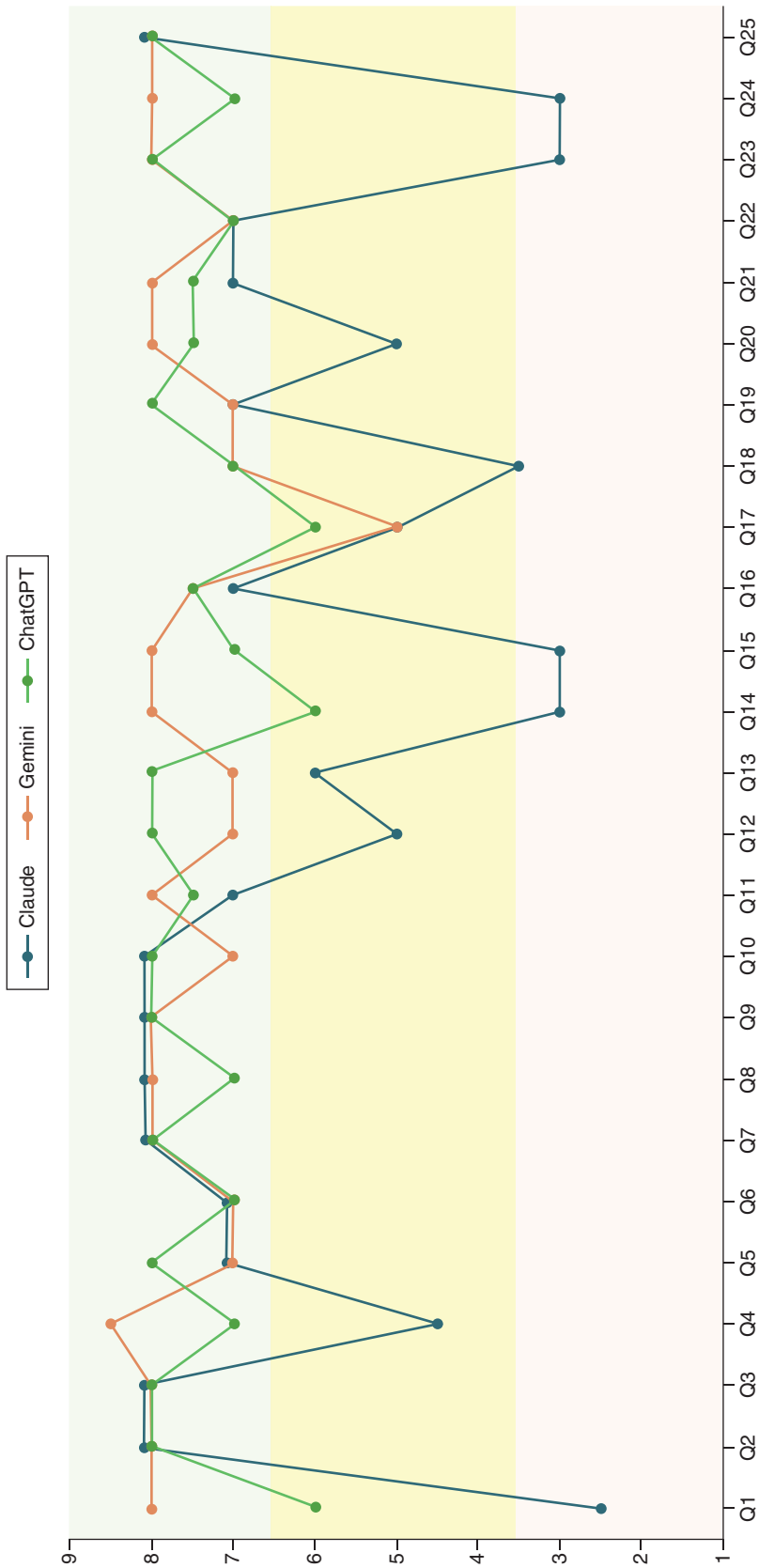


Figure 1. Median scores assigned by panelists for each question and large language model. Median scores on a 9-point scale are shown for each clinical question and large language model (Claude, Gemini and ChatGPT). Score ranges were defined as: 1–3 = disagreement, 4–6 = no consensus, 7–9 = agreement.

10.57264/cer-2026-0081

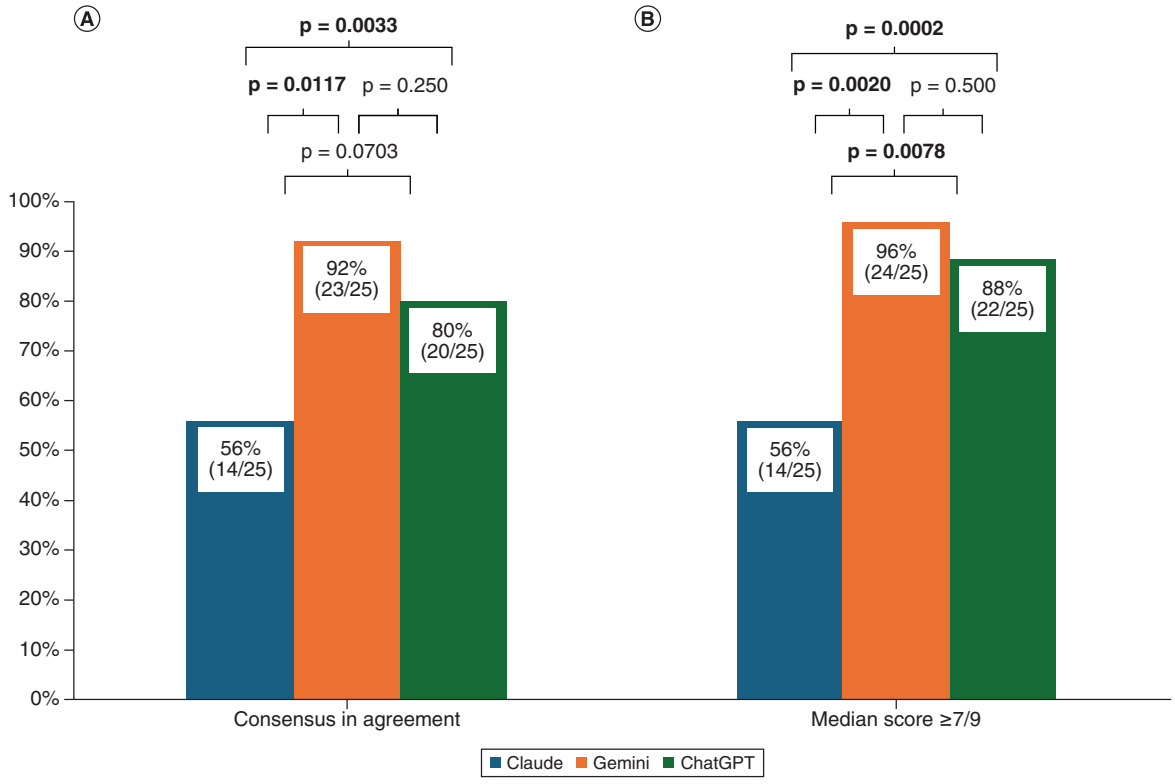


Figure 2. Comparison of expert panel agreement across large language models in prostate cancer. (A) Proportion of responses reaching consensus in agreement and **(B)** high-scoring items (median ≥7/9) across 25 responses for each large language model in prostate cancer.

highlight that although LLMs often deliver guideline-concordant recommendations in routine PC scenarios, their outputs can be variable in more challenging clinical contexts and should be interpreted with expert oversight. Importantly, this study did not assess LLMs as autonomous clinical decision-makers in real-world clinical practice.

LLMs are increasingly being explored as decision-support tools in oncology [15,16]. These AI systems are trained on vast corpora of biomedical literature and clinical guidelines, equipping them to generate detailed and context-specific responses to challenging clinical questions [15,16]. In PC care, LLMs and similar AI-driven platforms may help synthesize complex clinical information, support treatment recommendations, and facilitate the practical application of contemporary guidelines in everyday clinical settings [17–20]. Although these technologies show great potential to improve evidence-based decision-making, their clinical accuracy and reliability remain variable, especially in complex or evolving scenarios, highlighting the need for caution before implementation in patient care [15,21].

A closer look at our findings shows variability in median panel scores across the 25 questions. Standard management items generally received high scores, while questions on complex patients, advanced imaging, genetic testing, or those lacking sufficient clinical details were more likely to score below 7 for one or more models. Claude, in particular, recorded consensus in disagreement for three items (questions 15, 23 and 24), highlighting specific scenarios where its recommendations were considered inappropriate by most panelists. Qualitative feedback noted occasional omissions, limited clinical nuance and some factual inaccuracies, including clinically relevant issues in treatment sequencing, imaging interpretation, follow-up strategies and systemic treatment selection. These results indicate that, although LLMs can support clinical reasoning in PC, their use should remain subject to expert validation, especially in areas with limited or ambiguous guidance.

Recent evidence in PC supports a cautious yet optimistic perspective on integrating LLMs into clinical and educational practice. Riaz *et al.* described how AI, including LLMs, is expanding its role across the PC continuum, from early detection and risk stratification to individualized treatment selection and evidence synthesis for clinical guidelines [17]. They noted that these tools can integrate multimodal data and automate complex tasks, but

highlighted challenges related to data quality, model interpretability, and the necessity of human oversight [17]. In a comparative analysis, Geantă *et al.* evaluated three LLMs (ChatGPT 3.5, CoPilot and Gemini) against the official Patient's Guide using 25 expert-validated questions, with performance assessed for accuracy, timeliness, completeness and understandability through a Likert scale [18]. ChatGPT 3.5 achieved the highest overall ratings across all criteria, followed by CoPilot, while Gemini showed less consistent results. The authors observed that model performance consistently varied according to the type and complexity of the clinical context, indicating strengths in certain question types and weaker performance in others. They stressed that the safe implementation of these technologies in PC care will require robust ethical and regulatory frameworks, ongoing refinement and expert oversight. Importantly, the authors emphasized that future research should examine potential biases in AI-generated responses and their impact on patient outcomes, a recommendation with which we fully agree to ensure the equitable, safe and effective use of these technologies in PC care [18]. Extending this evidence, Trapp *et al.* assessed the responses of ChatGPT-4, Gemini, Copilot and Claude to common questions about radiotherapy for localized PC, incorporating evaluations from both clinicians and patients. While clinicians generally rated the responses as relevant and accurate, noting differences in completeness, clinical nuance and accuracy, patients found ChatGPT-4's answers clear, relevant and trustworthy, reporting increased confidence and willingness to use such tools for future medical queries [19]. Collectively, these studies indicate that although LLMs hold promise for enhancing decision-making and improving health literacy in PC, their integration into practice should proceed with continuous model refinement, external validation and expert supervision, particularly in complex or guideline-limited scenarios.

A notable challenge in evaluating LLMs in PC is the absence of a standardized methodology across studies. As noted by Carl *et al.*, current research applies a wide variety of evaluation strategies, including multiple-choice assessments, open-ended clinical prompts and comparisons with expert recommendations, making direct comparison across studies difficult [15]. In our study, we employed a Delphi process with a multidisciplinary expert panel from urology, medical oncology, radiation oncology, nuclear medicine, pathology, radiology and hospital pharmacy, providing the advantages of structured consensus-building, iterative review and rich qualitative feedback. Nevertheless, alternative evaluation frameworks may also yield valuable insights, and the existing heterogeneity in study designs highlights the importance of moving toward greater methodological harmonization in future research on the medical applications of LLMs in oncology.

This issue is particularly relevant when interpreting the differences observed between Claude, Gemini and ChatGPT in our current study. In previous research across major urological cancers, we used a direct expert rating approach with nine to eleven multidisciplinary uro-oncology specialists independently scoring each response on a standardized 5-point Likert scale, which suggested Claude's superior performance [8]. However, the present study focused exclusively on PC and applied a Delphi methodology involving iterative rounds of expert feedback and consensus building. Under this design, Gemini and ChatGPT, both based on versions launched in 2025, demonstrated higher levels of agreement among panelists, whereas Claude's performance, from a version launched in late 2024, was notably lower, with several items showing disagreement. Although methodological differences between studies contribute to these variances, a more likely cause is the rapid evolution and frequent updating of LLM architectures and training datasets, resulting in changes to model behavior over relatively short time frames. These observations highlight the necessity of ongoing validation and cautious interpretation of LLM performance, as results may not remain stable or comparable across different versions or models, even within the same model family.

These findings collectively emphasize the potential of LLMs as supportive tools in PC, especially in standard or resource-limited settings, where they can enable quick access to guideline-based recommendations and help reduce care disparities [15,22]. However, their adoption into practice should occur gradually, starting with low-risk tasks such as education, documentation or triage, and expanding only when strong evidence supports their safety and clinical usefulness [22,23]. Transparent reporting of model characteristics and independent validation remains crucial, along with close cooperation among medical experts, AI developers, and patients to ensure the model meets real-world needs [15,23]. To promote safe implementation, clinical guidelines, especially in complex fields, should use clear and structured language that can be understood by both humans and AI systems.

This study presents several limitations that should be considered when interpreting its findings. It focused exclusively on PC and included a limited set of 25 predefined clinical scenarios, which may not capture the full complexity and variability of real-world PC management and could limit the generalizability of the results to other tumor types or more diverse clinical settings. Additionally, the evaluation relied on a Delphi process involving a multidisciplinary panel of experts. While this method provides a structured approach for consensus building

and the collection of qualitative insights, it remains inherently subjective and may be influenced by the panel composition and individual interpretations. In some questions, especially those with greater specialty specificity, only a subset of the panel voted, and consensus was calculated exclusively among those who responded. While this increases content expertise for each item, it may introduce potential bias due to variable sample sizes across questions, which should be taken into account when interpreting the results. In addition, some lower-scoring items involved incomplete clinical scenarios or topics with limited or controversial guideline support, which may have made direct comparisons between models more difficult. Moreover, the rapid evolution and frequent updates of LLMs mean that their performance is dynamic and may differ significantly after the evaluation period. Despite these limitations, the study has important strengths. Specifically, guideline figures and tables were carefully reformulated into clear, structured language to reduce ambiguity for the AI models, improving the comparability of responses. Furthermore, employing a multidisciplinary Delphi methodology enriched the clinical relevance and validity of the evaluation by combining quantitative consensus with qualitative analysis of model outputs.

In conclusion, our results indicate that advanced LLMs, when provided with PC guidelines with plain-language adaptations of figures and tables, can generate responses that frequently align with expert consensus in routine clinical scenarios. Nonetheless, notable limitations persist in more complex or nuanced situations, where clinical expertise and judgment remain indispensable. While LLMs hold promise as decision-support tools in PC care, their outputs should be interpreted with caution and always under expert supervision. To facilitate their safe and effective integration into clinical workflows, we recommend that urological oncology guidelines adopt clear, standardized and structured formats that are readily interpretable by clinicians and AI systems, and that future research continues to refine evaluation methodologies, monitor for potential bias and ensure consistent performance as LLMs evolve.

Summary points

- Large language models may help interpret prostate cancer (PC) guideline information, but their outputs require expert review.
- This study evaluated Claude, Gemini and ChatGPT using 25 PC questions derived from European PC guideline content.
- The model responses were assessed by a multidisciplinary expert panel through a two-round Delphi process.
- Gemini and ChatGPT showed higher alignment with expert opinion than Claude.
- Consensus in agreement was reached for 23 of 25 Gemini responses, 20 of 25 ChatGPT responses and 14 of 25 Claude responses.
- Claude reached consensus in disagreement for three responses, highlighting clinically relevant limitations of model performance in some scenarios.
- Model performance was stronger for routine PC scenarios than for complex, incomplete or controversial clinical situations.
- The findings support the potential role of large language models as guideline-interpretation tools under expert supervision, rather than as autonomous clinical decision-makers.

Supplementary data

To view the supplementary data that accompany this paper please visit the journal website at https://becarispublishing.com/doi/epdf/10.57264/cer-2026-0081

Author contributions

A Borque-Fernando led the project and contributed to the study's conceptualization, methodology, project administration and supervision. All authors contributed jointly to the investigation, formal analysis, data curation, validation, resources, visualization, writing of the original draft and review and editing of the manuscript, and approved the final version for publication.

Acknowledgments

The authors thank Pablo Rivas for his assistance with medical writing on behalf of Content Ed Net.

Financial disclosure

This project was funded by the Spanish Association of Urology. The authors received support from Instituto de Investigación Sanitaria Aragón (Spain) for the project "Uro-Onco GuÍAs: Desarrollo de un Chatbot basado en IA Generativa para la Interpretación de Guías Clínicas en Uro-Oncología".

Competing interests disclosure

The authors have no financial and/or nonfinancial competing interests or relevant affiliations with any organization/entity to declare that are relevant to the subject matter or materials discussed in this manuscript. This includes employment, grants or research funding, consultancies, membership on scientific or other advisory boards, honoraria, stock ownership or options, paid expert testimony, patents received or pending or royalties.

Writing disclosure

Medical writing of the manuscript through Content Ed Net was funded by the Spanish Association of Urology.

Ethical conduct of research

This study involved only expert panelists who participated voluntarily in the Delphi process. No patient data were used. Participation implied informed consent, and formal ethical approval was not required.

Consent to participate

Participation by expert panelists was voluntary and implied informed consent.

Data sharing statement

The results of the Delphi process are presented within the manuscript. The complete set of responses generated by each large language model to all 25 questions are available for consultation at https://www.contentednet.com/proyecto-urooncoguias-aeu/

Open access

This work is licensed under the Attribution-NonCommercial-NoDerivatives 4.0 Unported License. To view a copy of this license, visit https://creativecommons.org/licenses/by-nc-nd/4.0/

References

Papers of special note have been highlighted as: • of interest; •• of considerable interest

1. Ferlay J, Ervik M, Lam F *et al.* Prostate cancer fact sheet. *Global Cancer Observatory: Cancer Today*. International Agency for Research on Cancer, Lyon, France (2024). (Accessed 5 August 2025). https://gco.iarc.who.int/media/globocan/factsheets/cancers/27-prostate-fact-sheet.pdf
2. EAU Guidelines on Prostate Cancer - Uroweb. (Accessed 19 May 2025). https://uroweb.org/guidelines/prostate-cancer
3. Cai T, Cocci A, Gacci M *et al.* Guidelines in urology: lights and shadows. *Urologia* 87(3), 125–129 (2020).
4. Smith EJ, MacLennan S, Bjartell A *et al.* Ensuring consistent European-wide urological care by the use of evidence-based clinical practice guidelines: can we do better. *Biomed. Hub.* 2, 162–168 (2017).
5. Shore N, Brown G, Lowentritt B *et al.* Evaluating the importance of practice-level factors on adherence to prostate cancer treatment guidelines in urology. *Urology* 196, 147–154 (2025).
6. Benary M, Wang XD, Schmidt M *et al.* Leveraging large language models for decision support in personalized oncology. *JAMA Netw. Open* 6, e2343689 (2023).
7. Kaiser P, Yang S, Bach M *et al.* The interaction of structured data using open EHR and large language models for clinical decision support in prostate cancer. *World J. Urol.* 43, 67 (2025).
8. Borque-Fernando Á, Navarro D, Doblare M *et al.* Accuracy of large language models in interpreting urological clinical guidelines: a comparative study with expert evaluation. *Ther. Adv. Urol.* 18, 17562872261436905 (2026).

•• **Evaluates Claude, Gemini and ChatGPT across major urological cancers using guideline-based questions, plain-language guideline adaptation and multidisciplinary expert assessment.**

9. Meet Claude. Anthropic. (Accessed 23 June 2025). https://www.anthropic.com/claude
10. Google Gemini. Gemini. (Accessed 23 June 2025). https://gemini.google.com
11. GPT-4o mini Model | OpenAI API. OpenAI. (Accessed 23 June 2025). https://developers.openai.com/api/docs/models/gpt-4o-mini
12. Models overview. Anthropic. (Accessed 23 June 2025). https://platform.claude.com/docs/en/about-claude/models/overview
13. Gemini 2.5 Flash | Gemini API. Google. (Accessed 23 June 2025). https://ai.google.dev/gemini-api/docs/models/gemini-2.5-flash?hl=es-419
14. GPT-4.1 mini Model | OpenAI API. OpenAI. (Accessed 23 June 2025). https://developers.openai.com/api/docs/models/gpt-4.1-mini
15. Carl N, Schramm F, Haggenmüller S *et al.* Large language model use in clinical oncology. *NPJ Precis. Oncol.* 8, 240 (2024).

•• **Provides a broad overview of current applications, methodologies and performance of large language models (LLMs) in clinical oncology, while highlighting heterogeneity in evaluation methods and the need for standardized reporting.**

16. Zitu MM, Le TD, Duong T *et al.* Large language models in cancer: potentials, risks, and safeguards. *BJR Artif. Intell.* 2, ubae019 (2025).

17. Riaz IB, Harmon S, Chen Z, Naqvi SAA, Cheng L. Applications of artificial intelligence in prostate cancer care: a path to enhanced efficiency and outcomes. *Am. Soc. Clin. Oncol. Educ. Book* 44, e438516 (2024).

• **Summarizes AI applications across the prostate cancer care pathway and discusses the role of LLMs in information summarization, evidence synthesis, clinical trial design, drug development and clinical practice guidelines.**

18. Geantă M, Bădescu D, Chirca N *et al.* The emerging role of large language models in improving prostate cancer literacy. *Bioengineering (Basel)* 11, 654 (2024).

• **Compares LLM-generated prostate cancer information with an official patient guide, assessing accuracy, timeliness, completeness and understandability.**

19. Trapp C, Schmidt-Hegemann N, Keilholz M *et al.* Patient- and clinician-based evaluation of large language models for patient education in prostate cancer radiotherapy. *Strahlenther. Onkol.* 201, 333–342 (2025).

• **Assesses LLM-generated responses to common patient questions about localized prostate cancer radiotherapy, including clinician ratings of relevance, correctness and completeness, as well as patient perceptions of comprehensibility and trustworthiness.**

20. Khattak A, Ruta D, Patkar V *et al.* Artificial intelligence-driven streamlining of prostate cancer multidisciplinary team recommendations in a tertiary NHS centre in the UK. *BJU Int.* 136(5), 872–881 (2025).

21. Hager P, Jungmann F, Holland R *et al.* Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nat. Med.* 30, 2613–2622 (2024).

•• **Evaluates LLM limitations in realistic clinical decision-making scenarios and shows that current models are not ready for autonomous clinical use.**

22. Longwell JB, Hirsch I, Binder F *et al.* Performance of large language models on medical oncology examination questions. *JAMA Netw. Open* 7, e2417641 (2024).

23. Ah-Thiane L, Heudel PE, Campone M *et al.* Large language models as decision-making tools in oncology: comparing artificial intelligence suggestions and expert recommendations. *JCO Clin. Cancer Inform.* 9, e2400230 (2025).