THERAPEUTIC ADVANCES in
Urology

Original Research

Accuracy of large language models in interpreting urological clinical guidelines: a comparative study with expert evaluation

Ther Adv Urol

2026, Vol. 18: 1–17

DOI: 10.1177/17562872261436905

Ángel Borque-Fernando, Denis Navarro, Manuel Doblare, Luis M. Esteban, Daniel Perez-Fentes, Mario Álvarez-Maestro, Rafael A. Medina-López, Oscar Rodríguez Faba, José Rubio-Briones, Sergio Fernández-Pello, Jesús María Fernández-Gómez, Tomás Fernández Aparicio, Félix Guerrero Ramos, Laura Izquierdo and José Luis Álvarez-Ossorio Fernández

Abstract

Background: Large language models (LLMs) are increasingly being explored to supporting evidence-based decision-making in urology, but their accuracy in interpreting and applying clinical guidelines remains uncertain.

Objectives: We aimed to evaluate the ability of LLMs to interpret and apply clinical guidelines across the full spectrum of major urological cancers.

Design: This expert-validated study evaluated six configurations of three top LLMs (Claude, Gemini, and ChatGPT) using 25 structured questions for each of the seven major urological cancers: prostate cancer, upper tract urothelial carcinoma, muscle-invasive and non-muscle-invasive bladder cancer, renal cell carcinoma, penile cancer, and testicular cancer.

Methods: Both simple and rephrased prompts were used to assess the impact of prompt engineering on response quality. All figures and tables from the English-language EAU guidelines were systematically converted into plain, structured text and peer reviewed by multidisciplinary experts before evaluating the LLM responses. Each response was independently rated by 9–11 uro-oncology specialists using a five-point Likert scale (1: incorrect/unacceptable, 5: optimal), resulting in 10,500 evaluations.

Results: Claude achieved the highest overall accuracy, with 45.9% of responses rated as optimal (Likert 5) and 87% as optimal/acceptable (Likert 4–5). Tumor-specific performance peaked in muscle-invasive bladder (56.7% optimal, 93% optimal/acceptable), penile (49.5%, 95%), and testicular cancer (60.9%, 94%). Gemini and ChatGPT showed lower optimal rates but acceptable performance (68%–70% optimal/acceptable). Rephrased prompts did not consistently outperform simple versions. All models showed acceptable accuracy, but the results should be interpreted cautiously due to recency bias and fast LLM tech evolution.

Conclusion: This study demonstrates the value of rigorous plain language adaptation and expert validation in benchmarking LLMs, supporting their potential as decision-support tools in uro-oncology.

Keywords: artificial intelligence, clinical, decision support systems, large language models, natural language processing, practice guidelines as topic, urologic neoplasms, validation study

Received: 26 August 2025; revised manuscript accepted: 27 February 2026.

Correspondence to:
Ángel Borque-Fernando
Unidad de Próstata, Servicio de Urología, Hospital Universitario Miguel Servet, Zaragoza, Spain

Área de Urología, Departamento de Cirugía, Facultad de Medicina, Universidad de Zaragoza, Zaragoza, Spain

Grupo de Investigación "URO-SERVET," IIS-Aragón, Zaragoza, Spain

Grupo de Ingeniería y Ciencia de Datos Aplicada, Escuela Universitaria Politécnica de La Almunia, Universidad de Zaragoza, Zaragoza, Spain
dr.borque@gmail.com

Denis Navarro
Department of Electronic Engineering and Communications I3A, Universidad de Zaragoza, Zaragoza, Spain

Manuel Doblare
Aragón Institute of Engineering Research (I3A), Zaragoza, Spain

Aragón Institute of Health Research (IIS Aragón), Zaragoza, Spain

Luis M. Esteban
Escuela Universitaria Politécnica de La Almunia, Instituto de Biocomputacion y Física de Sistemas Complejos, Universidad de Zaragoza, Zaragoza, Spain

Daniel Perez-Fentes
Urology Department, Santiago de Compostela University Hospital Complex, Santiago de Compostela, Spain

Mario Álvarez-Maestro
Servicio Urología, Hospital Universitario La Paz, Madrid, Spain

Rafael A. Medina-López
Servicio Urología, Hospital Universitario Virgen del Rocío, Instituto de Biomedicina de Sevilla (IBiS), Sevilla, Spain

Oscar Rodríguez Faba
Servicio de Urología, Fundació Puigvert Barcelona, Barcelona, Spain

José Rubio-Briones
Servicio de Urología, Hospital VITHAS 9 de Octubre, Valencia, Spain

Sergio Fernández-Pello
Servicio de Urología, Hospital Universitario de Cabueñes, Gijón, Spain

Jesús María Fernández-Gómez
Hospital Universitario Central de Asturias, Universidad de Oviedo, Oviedo, Spain

Tomás Fernández Aparicio
Servicio de Urologia, Hospital Morales Meseguer, Murcia, Spain

Félix Guerrero Ramos
Unidad de UroOncología, Servicio de Urología, Hospital Universitario 12 de Octubre, Madrid, Spain

Unidad de Urooncología, Departamento de Urología, Hospital Universitario HM Sanchinarro, Hospital Universitario HM Montepríncipe, Hospital Universitario HM Puerta del Sur, Madrid, España

Facultad de Medicina, Universidad San Pablo CEU, Madrid, Spain

Laura Izquierdo
Department of Urology, Hospital Clinic, Barcelona, Spain

Genetics and Urologic Tumors, Institut d'Investigacions Biomèdiques August Pi i Sunnyer (IDIBAPS), Barcelona, Spain

Department of Surgery and Medical-Surgical Specialties, Medicine and Health Sciences Faculty, Universitat de Barcelona, Barcelona, Spain

José Luis Álvarez-Ossorio Fernández
Director del Patronato de la Fundación para la Investigación en Urología (Asociación Española de Urología), Jefe de Servicio Urología Hospital Universitario Puerta del Mar, Cádiz, Spain

Introduction

Large language models (LLMs) such as GPT-4o,[1] Gemini,[2] and Claude[3] have rapidly emerged as transformative tools in healthcare, offering clinicians and researchers unprecedented access to up-to-date medical knowledge and the ability to query complex clinical guidelines in natural language.[4,5] In urology, where clinical decision-making is increasingly guided by detailed and frequently updated recommendations from organizations such as the European Association of Urology (EAU),[6] the potential of LLMs to interpret, summarize, and contextualize guideline content is especially promising. However, despite their impressive capabilities, the reliability and accuracy of LLM-generated responses to guideline-based queries have not been fully assessed, particularly when nuanced clinical judgment or the integration of multidisciplinary evidence is required. This has led to a growing need for systematic evaluation of LLM performance against expert consensus, with the aim of identifying which models provide the most trustworthy and clinically relevant answers to urological guideline questions.[7–9]

LLMs face significant challenges when applied to interpreting clinical guidelines in urology. Recent studies have demonstrated notable variability in the accuracy, consistency, and completeness of LLM-generated responses, with performance differing not only between models but also depending on the complexity of the clinical scenario and the phrasing of queries. For example, while some models, such as GPT-4o and Gemini 1.5 Pro, achieve accuracy rates approaching 80% in guideline-based tasks when optimized with expert-driven prompts, others still struggle with nuanced questions, negative phrasing, or require domain-specific training to minimize errors and hallucinations.[10] These limitations highlight the ongoing need for rigorous expert evaluation and standardized benchmarks to ensure that LLM outputs are both clinically reliable and aligned with current best practices in urology.

In this study, we aimed to evaluate the ability of LLMs to interpret and apply clinical guidelines across the full spectrum of major urological cancers, including prostate cancer,[11] upper tract urothelial carcinoma,[12] non-muscle-invasive and muscle-invasive bladder cancer,[13,14] testicular cancer,[15] renal cell carcinoma,[16] and penile cancer.[17] To accomplish this, we implemented six LLM configurations using the latest low-cost versions of GPT-4o-mini-2024-07-18,[1] Claude-Haiki-2024-10-22,[3] and Gemini-Flash-001 (May 2024).[2] Each model was assessed using two distinct approaches: a simple method with direct queries and a rephrased method that optimized prompts for greater clinical relevance. For each cancer type, model responses to a set of 25 expert-validated, guideline-based questions were independently rated by a multidisciplinary panel of uro-oncology experts using a standardized Likert scale. This design enabled a quantitative comparison of clinical accuracy and usability across models and tumor types.

Methods

Study design and objectives

A comparative evaluation of LLMs in uro-oncology was conducted to identify the model that delivers the highest clinical accuracy and usability for guideline-based clinical questions, based on experts' opinions. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) guidelines were consulted when preparing the manuscript, according to the EQUATOR Network reporting guidelines, and the completed checklist has been provided as Supplemental Material.

Expert panel composition

The project involved a multidisciplinary group of scientific and technical experts from leading Spanish scientific societies, including the following: (1) Scientific board: composed of president and directors of the Spanish Association of Urology (AEU), head of urological research institute, and coordinators of uro-oncological groups, representing all major cancer subtypes; (2) Technical board: formed by university professors specializing in biomedical and electronics engineering, principal investigators in AI and digital health, and a computer engineer from the AEU providing informatics support; (3) Specialized tumor boards: consisting of seven boards, each dedicated to a specific malignancy (e.g., muscle-invasive bladder cancer, upper tract tumors). Each board was led by a renowned expert in the field and included urologists, radio-oncologists, medical oncologists, nuclear medicine physicians, pathologists, pharmacologists, and radiologists from the 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), respectively.

Guidelines adaptation and questionnaire development

Before developing questions, all figures and tables from the clinical guidelines were systematically adapted into plain, structured textual descriptions to facilitate accurate interpretation by LLMs. To ensure the fidelity of this adaptation, each guideline was divided into thematic sections that underwent collaborative peer review: independent reviewers from the respective tumor boards examined each section, certifying the accuracy of the adaptation or providing corrections. Guideline coordinators resolved any discrepancies through consensus, with final approval requiring unanimous validation.

Building on this rigorously validated adaptation process for the guidelines, a set of 25 structured questions was specifically designed to assess the ability of LLM to interpret and apply clinical guidelines in uro-oncology. Each set of questions was developed to evaluate one of the 2024 EAU guidelines, specifically covering prostate cancer,[11] upper urinary tract urothelial carcinoma,[12] muscle-invasive and non-muscle-invasive bladder cancer,[13,14] renal cell carcinoma,[16] penile cancer,[17] and testicular cancer.[15] Authorization was provided by the EAU for the use of the guidelines in this project. The whole question set is available in the Supplemental Material (Tables S1–S7). These questions comprehensively addressed key clinical domains, including diagnosis, risk stratification, indications for diagnostic testing, and therapeutic management. Each questionnaire underwent further validation by a multidisciplinary panel of 70 experts from AEU, SEOR, GUARD, SEMNIM, SEAP, SEFH, and SERAM, ensuring clinical pertinence and alignment with current best practices across all relevant specialties.

LLM selection and model assessment

Six publicly accessible LLMs were evaluated, corresponding to the latest low-cost versions of three platforms: GPT-4o-mini-2024-07-18 (OpenAI),[1] Gemini-Flash-001 (Google, May 2024),[2] and Claude-Haiki-2024-10-22 (Anthropic),[3] each tested using both a simple and a rephrased prompt approach. These models were selected based on their availability and affordability at the time of the study. The LLM responses were generated exclusively based on adapted plain-language guidelines content, which served as the sole information source. This approach ensures that the models' performance reflects their ability to interpret and apply guideline-based information without influence from external sources or general training data, thereby providing a focused and clinically relevant evaluation.

To evaluate the effect of prompt design on response quality, each model was assessed using two distinct approaches: simple and rephrased. The simple approach involved directly querying the model with the original question, employing standard prompt engineering techniques. By contrast, the rephrased approach used a two-step process: first, the model was prompted to reformulate the original question by integrating relevant clinical context and targeting expert-level precision; next, the enhanced question was submitted back to the model using the same direct querying method. For each of the three LLMs, distinct prompts were specifically optimized using each respective model. This tailored approach ensures that performance comparisons across models accurately reflect the intrinsic capabilities of each LLM rather than differences resulting solely from generic prompting techniques. Throughout this process, iterative refinement of prompts was conducted to improve question clarity and clinical specificity, allowing a systematic assessment of how prompt structure influences the accuracy and relevance of model responses, in line with current recommendations in medical prompt engineering.[18–20] An example of the simple and rephrased prompt design is provided in the Supplemental Materials (Table S8).

Assessment procedure

Each LLM response was independently scored by 9–11 uro-oncology experts from the multidisciplinary panel described above. Evaluators received standardized instructions and accessed responses via a digital platform, rating answers on a five-point Likert scale (1: Incorrect/unacceptable; 5: Optimal). Qualitative feedback was permitted to provide context for ratings. The evaluation was conducted in Spanish because it is the native language of the expert board that reviewed the questionnaire. This process of simultaneous double

translation does not undermine the accuracy of the required information.

Two primary dimensions were subjectively assessed by the experts from each answer: (1) Clinical accuracy: Alignment with current evidence and guidelines; (2) Usability: Clarity, relevance, and practicality for clinical decision-making. These two aspects were not scored separately but were considered together in each rating, ensuring that only responses that were both clinically correct and practically useful received high scores.

To minimize evaluator fatigue and bias, each assessment session was limited to a maximum of five questions, corresponding to approximately 1 h per session. Evaluators were allowed to abstain from rating any question outside their area of expertise by marking it as "N/A" (not applicable). All evaluation procedures and protocols were thoroughly documented to ensure transparency throughout the study.

Importantly, within our evaluation, we deliberately included a small subset of questions for which no answer was available from the Guidelines. Under our constrained setting ("answer strictly based on the provided guideline context"), models could either abstain (i.e., state that the answer could not be given from the supplied context) or fabricate content. Expert evaluators then expressed their level of agreement with the model's behavior—rewarding appropriate abstention and penalizing fabrication.

Statistical analysis

All evaluator scores were automatically recorded for each model and question. Following data collection, the results were summarized in comparative tables that displayed the absolute frequencies of responses for each model and guideline. The distribution of each Likert scale category (1–5) was shown with mosaic plots. Line charts were created to display the mean and standard deviation of Likert scale responses for each model across different guidelines. In addition, inter-rater agreement for each guideline was assessed using correlation heatmaps based on Spearman's rank correlation coefficient.

Future validation phase

To ensure the clinical applicability and robustness of the results, a future validation phase has been planned to follow the selection of the best-performing LLM. This phase will employ a structured Delphi study to rigorously assess the performance of the top LLMs in uro-oncology. For each of the seven specialized tumor boards, 25 new guideline-based questions will be generated and answered by the two leading LLMs identified during the initial evaluation.

These questions, along with their AI-generated responses, will be hosted on a secure, responsive web platform accessible from any device. Expert panelists will be invited to rate their level of agreement with each LLM response using a nine-point Likert scale, where one indicates strong disagreement and nine indicates strong agreement. The Delphi process will consist of two rounds: in the first round, panelists will provide their initial ratings; in the second round, questions that do not reach consensus will be re-evaluated after panelists receive anonymized feedback.

This approach is designed to facilitate consensus-building among multidisciplinary clinical experts, enabling a robust validation of LLM outputs against expert judgment and supporting the identification of the most clinically reliable and useful AI model for decision support in uro-oncology.

Results

A total of 10,500 individual expert evaluations were collected in this phase, corresponding to 25 questions per guideline, 6 model responses per question, a mean of 10 expert evaluators per guideline, and 7 uro-oncology guidelines. Each model was thus assessed in 1750 independent evaluations.

Global results across all guidelines

Claude simple obtained the highest number of optimal ratings, with 652 (45.9%) responses receiving a rating of 5. This was followed by Claude rephrased (565, 40.1%), Gemini rephrased (417, 29.6%), Gemini simple (331, 23.3%), ChatGPT simple (315, 22.2%), and ChatGPT rephrased (265, 18.8%). For responses rated as 4, Gemini simple (657, 46.4%) and ChatGPT simple (650, 45.8%) had the highest counts, followed by Claude simple (579, 40.8%) and Claude rephrased (541, 38.4%). The remaining models, Gemini rephrased (533, 37.8%) and ChatGPT rephrased (449, 31.8%), had lower counts in this category. Lower Likert scores (1–3) were less frequent across all models (Figures 1–3).

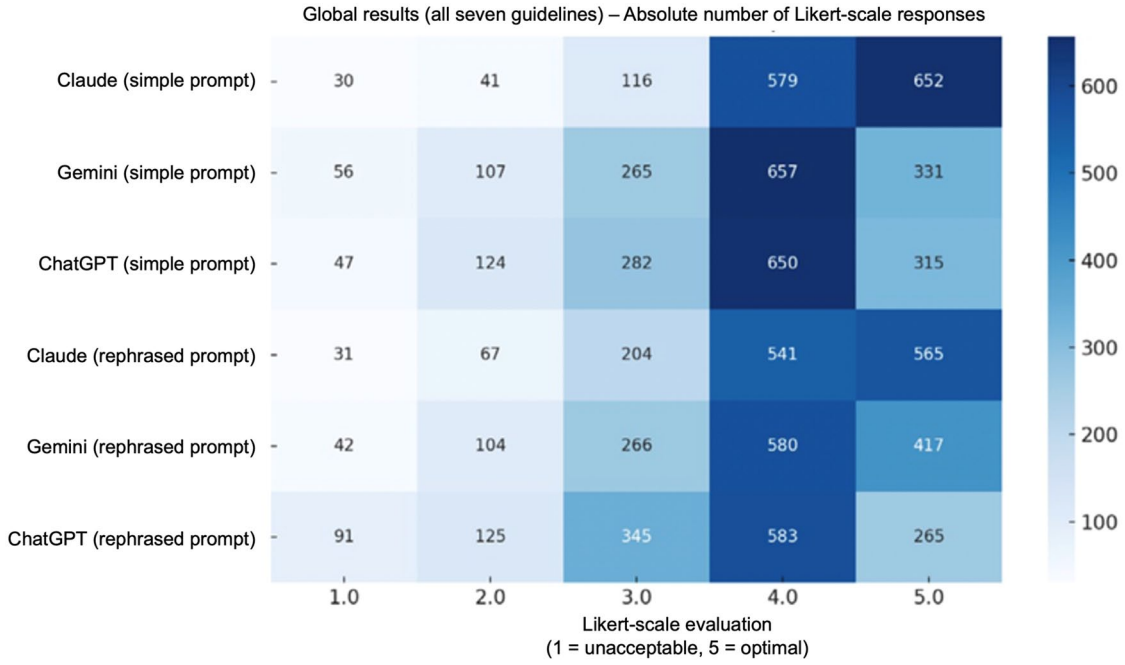


Figure 1. Absolute number of Likert-scale ratings for LLM-generated, guideline-based answers in uro-oncology across all seven guidelines, by model and prompt type.
LLM, large language model.

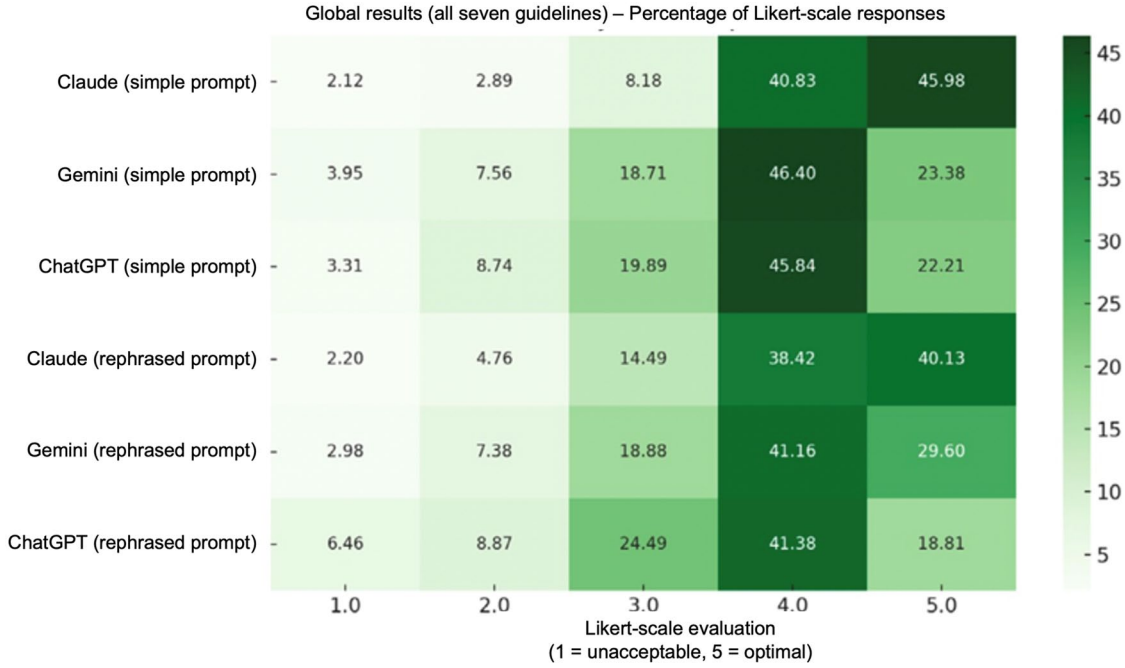


Figure 2. Percentage distribution of Likert-scale ratings for LLM-generated, guideline-based answers in uro-oncology across all seven guidelines, by model and prompt type.
LLM, large language model.

When considering both optimal and acceptable responses (scores 4 or 5), Claude simple prompt reached 87%, Claude rephrased prompt 78%, Gemini rephrased prompt 71%, Gemini simple prompt 70%, ChatGPT simple prompt 68%, and ChatGPT rephrased prompt 60%. All models

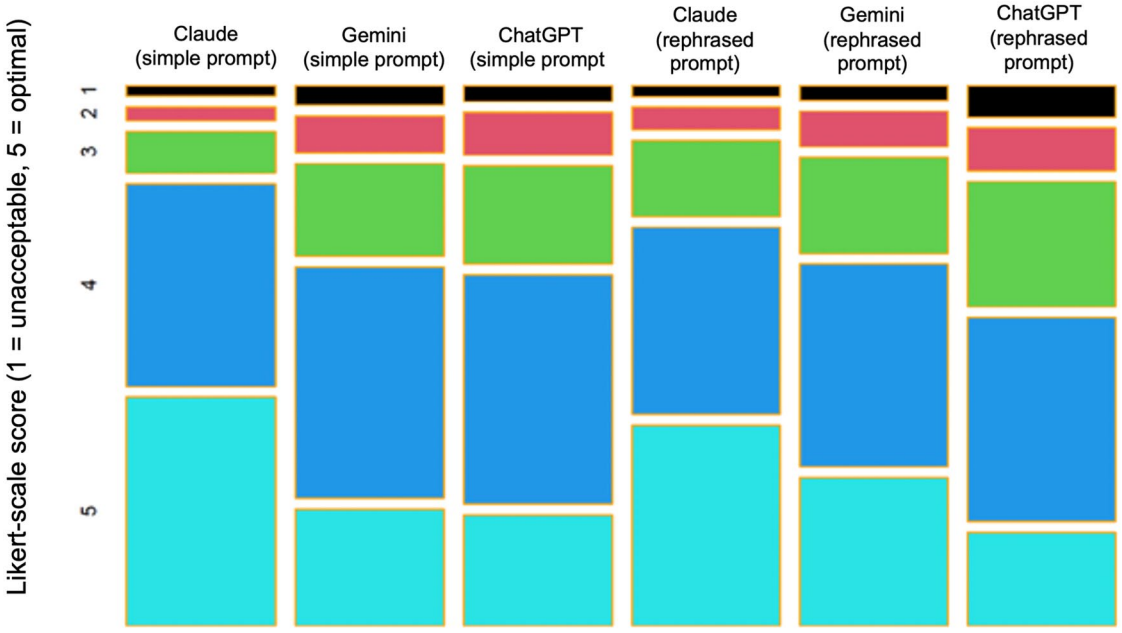


Figure 3. Illustrative percentage breakdown of Likert-scale ratings for LLM-generated, guideline-based answers in uro-oncology (all seven guidelines).
LLM, large language model.

Table 1. Cumulative percentage of Likert-scale responses by model and prompt type.

Likert-scale score (1 = unacceptable, 5 = optimal)	Claude (simple prompt) (%)	Gemini (simple prompt) (%)	ChatGPT (simple prompt) (%)	Claude (rephrased prompt) (%)	Gemini (rephrased prompt) (%)	ChatGPT (rephrased prompt) (%)
5	46	23	22	40	30	19
4	87	70	68	78	71	60
3	95	89	88	93	90	85
2	98	96	97	98	97	94
1	100	100	100	100	100	100

had more than 85% of responses in the top three Likert categories (scores 3–5; Table 1).

Detailed proportions of responses for each of the cancers are provided in the Supplemental Material (Figures S8–S21).

The mean Likert scores for each model were as follows: Claude's simple model had the highest mean, exceeding 4.2. Claude also rephrased and had a mean score above 4.0. Both Gemini simple and ChatGPT simple had mean scores below 3.9. The lowest mean scores were observed for Gemini rephrased and ChatGPT rephrased, with ChatGPT rephrased just above 3.6 (Figure 4). The median response scores for each type of cancer are summarized in the Supplemental Material (Figure S22).

Prostate cancer

For prostate cancer, Claude simple achieved the highest performance, with 49.6% of answers rated as acceptable (Likert 4) and 28.4% as optimal (Likert 5; 78.0% combined). Claude rephrased followed with 38.2% acceptable and 36.6% optimal (74.8% combined), while Gemini rephrased also performed strongly, with 40.6% acceptable and 31.7% optimal responses (72.3%

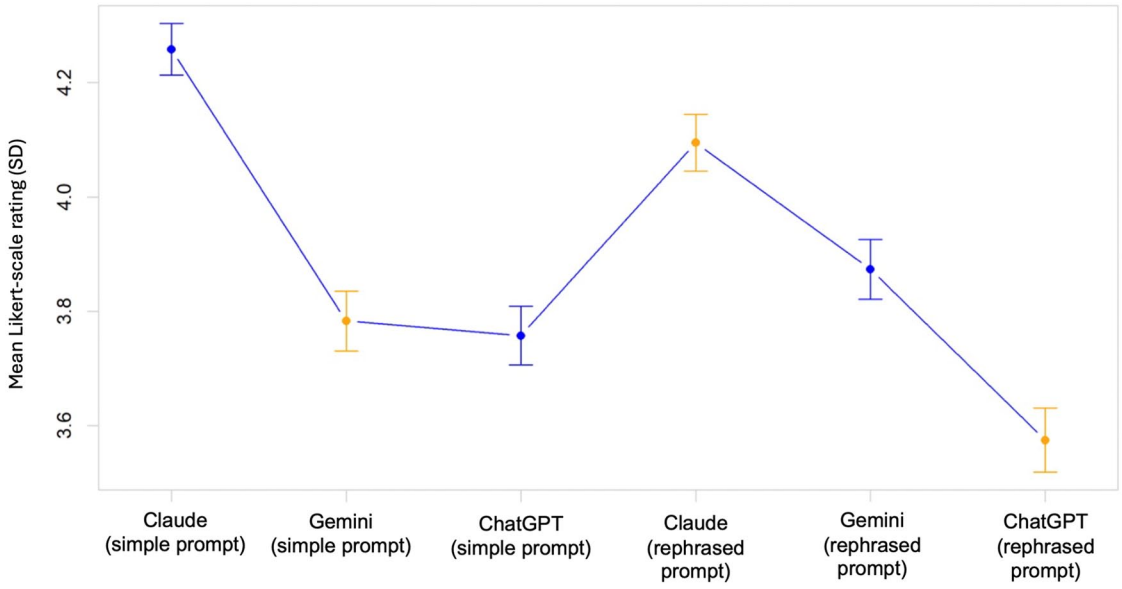


Figure 4. Mean (SD) Likert-scale ratings for each LLM and prompt type, aggregated across all seven uro-oncology guidelines. Error bars represent standard deviation.
LLM, large language model.

combined). The remaining models showed lower proportions of clinically valid responses (Figures S8 and S9).

Regarding the average Likert scores, Claude rephrased achieved the highest mean, approaching 4.0, while both Claude simple and Gemini rephrased also performed robustly, with mean scores close to 3.9. By contrast, the remaining models recorded lower average scores, between about 3.5 and 3.6 (Figure S22).

Upper urinary tract urothelial carcinoma

For upper urinary tract urothelial carcinoma, Claude simple achieved the strongest performance, with 52.7% of responses rated as optimal and 31.4% as acceptable, totaling 84.1% in the top two categories. Gemini simple ranked second, with 23.8% optimal and 41.3% acceptable responses (65.1% combined), followed by ChatGPT simple, which reached 15.9% optimal and 41.3% acceptable (57.2% combined). The rephrased versions of all models showed lower proportions of optimal and acceptable responses than their simple counterparts (Figures S10 and S11).

Claude simple achieved the highest mean Likert score, exceeding 4.2. Claude rephrased followed, with a mean near 3.8, while Gemini simple and ChatGPT simple both scored around 3.6. The lowest averages were observed for Gemini rephrased, around 3.4, and ChatGPT rephrased, just above 3.3 (Figure S22).

Muscle-invasive bladder cancer

For muscle-invasive bladder cancer, Claude simple demonstrated the strongest performance, with 56.7% of responses rated as optimal and 36.2% as acceptable, amounting to a total of 92.9% in the top two categories. Claude rephrased ranked second, with 47.6% optimal and 33.9% acceptable responses, totaling 81.5%. ChatGPT simple was the third-best performer, reaching 24.8% optimal and 49.5% acceptable responses (74.3% combined). The remaining models showed lower proportions of optimal or acceptable answers (Figures S12 and S13).

Claude simple obtained the highest mean Likert score, with an average above 4.4. Claude rephrased also performed well, with a mean score close to 4.2. Gemini simple, Gemini rephrased, and ChatGPT simple showed lower mean scores, around 3.8–3.9, while ChatGPT rephrased recorded the lowest mean, close to 3.5 (Figure S22).

Non-muscle-invasive bladder cancer

For non-muscle-invasive bladder cancer, Claude simple achieved the best performance, with

44.6% of responses rated as optimal and 40.1% as acceptable, totaling 84.7% in these two categories. Claude rephrased followed, with 40.3% of responses rated as optimal and 36.9% as acceptable (77.2% combined). The other models achieved a lower combined percentage of optimal and acceptable responses compared to the Claude models (Figures S14 and S15).

For muscle-invasive bladder cancer, Claude simple achieved the highest mean Likert score, with an average above 4.2. Claude rephrased followed by a mean close to 4.0. The remaining models showed lower average scores, all falling between 3.6 and 3.8 (Figure S22).

Renal cell carcinoma

For renal cell carcinoma, Claude simple achieved the highest performance, with 28.1% of responses rated as optimal and 48.5% as acceptable, totaling 76.6% in the top two categories. Claude rephrased followed with 29.5% optimal and 43.4% acceptable responses (72.9% combined), and Gemini simple ranked third with 12.0% optimal and 55.7% acceptable answers (67.7% combined). The remaining models yielded lower proportions of optimal or acceptable responses; however, the majority of answers across all models were still concentrated in the higher Likert categories (Figures S16 and S17).

Claude rephrased achieved the highest mean Likert score, with an average close to 4.1. Claude simple also performed strongly, with a mean above 4.0, and Gemini rephrased ranked third, just below 3.9. By contrast, Gemini simple and ChatGPT simple both averaged around 3.6, while ChatGPT rephrased recorded the lowest mean, slightly above 3.4 (Figure S22).

Penile cancer

For penile cancer, Claude simple achieved the best performance, with 49.5% of responses rated as optimal and 45.9% as acceptable, totaling 95.4% in the top two categories. Gemini simple ranked second, with 40.2% optimal and 42.3% acceptable responses (82.5% combined), followed by Claude rephrased with 32.7% optimal and 46.4% acceptable (79.1% combined). The remaining models had consistently lower combined proportions of optimal and acceptable responses, all staying below 75% (Figures S18 and S19).

Claude simple achieved the highest mean Likert score, with an average above 4.4. Gemini simple ranked second, with a mean above 4.1, and Claude rephrased followed closely, just over 4.0. The remaining models showed lower mean scores: ChatGPT simple was close to 3.9, Gemini rephrased slightly below 3.9, and ChatGPT rephrased had the lowest mean, near 3.4 (Figure S22).

Testicular cancer

For testicular cancer, Claude simple achieved the highest performance, with 60.9% of responses rated as optimal and 33.8% as acceptable, totaling 94.7% in the top two categories. Claude rephrased followed closely with 60.5% optimal and 32.7% acceptable responses (93.2% combined), and Gemini rephrased ranked third with 49.6% optimal and 35.7% acceptable responses (85.3% combined). By contrast, ChatGPT simple, Gemini simple, and ChatGPT rephrased showed lower proportions of optimal responses and did not reach the combined performance levels of the top three models, despite moderate rates of acceptable responses (Figures S20 and S21).

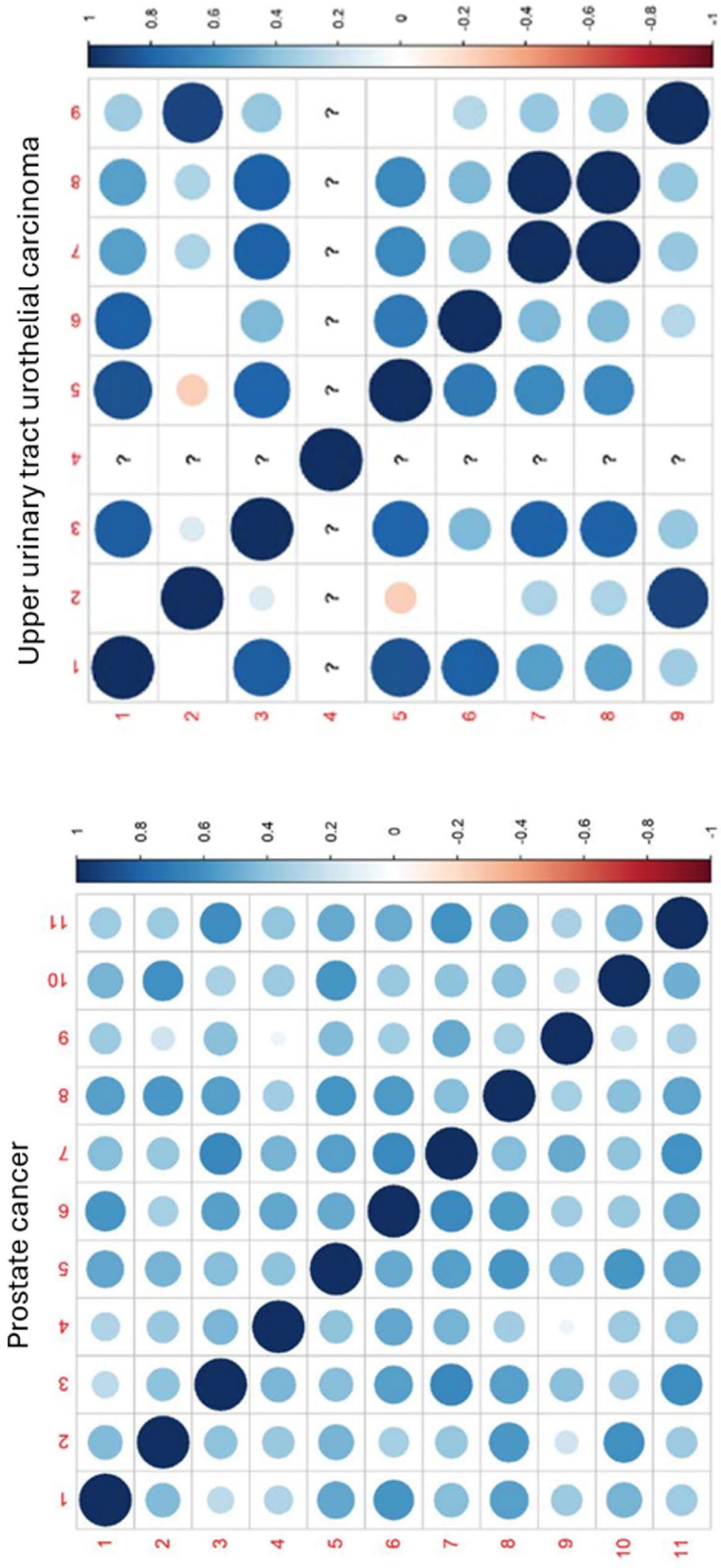
Claude simple achieved the highest mean Likert score, with an average above 4.5. Claude rephrased also performed strongly, with a mean close to 4.5, and Gemini rephrased ranked third, just below 4.3. The remaining models all had mean scores around 4.1 (Figure S22).

Inter-rater correlations

Regarding inter-rater correlation across the respective guidelines, the correlation heatmaps reveal a moderate level of agreement, with average Spearman correlation coefficients of 0.36, 0.28, 0.33, 0.30, 0.25, 0.31, and 0.25, respectively. These values indicate variability in the experts' responses (Figures 5–7).

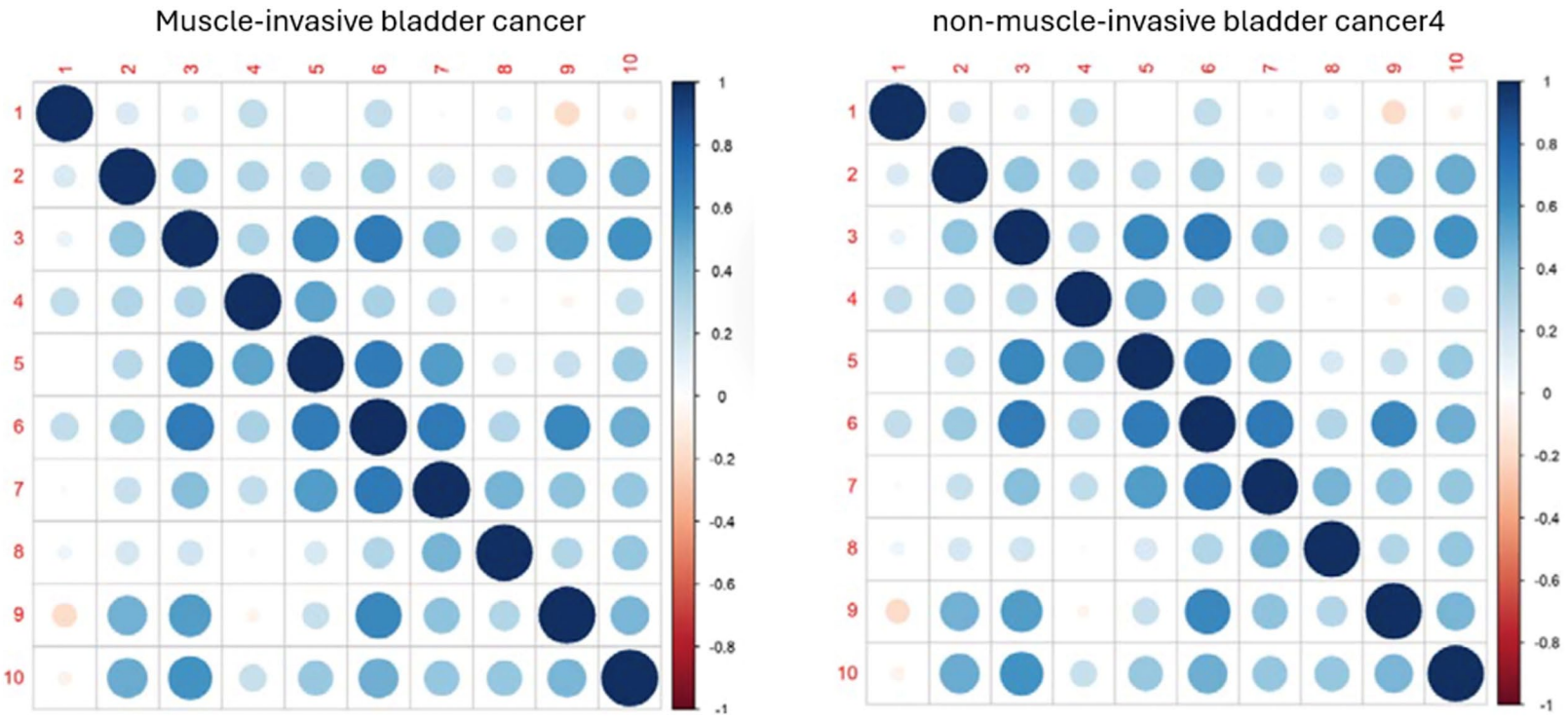
Discussion

This study provides a detailed comparative evaluation of LLMs in interpreting and responding to guideline-based questions in uro-oncology. By systematically assessing 10,500 answers generated by six different LLM configurations, including both simple and rephrased versions of Claude, Gemini, and ChatGPT, against 25 expert-validated questions covering the main urologic



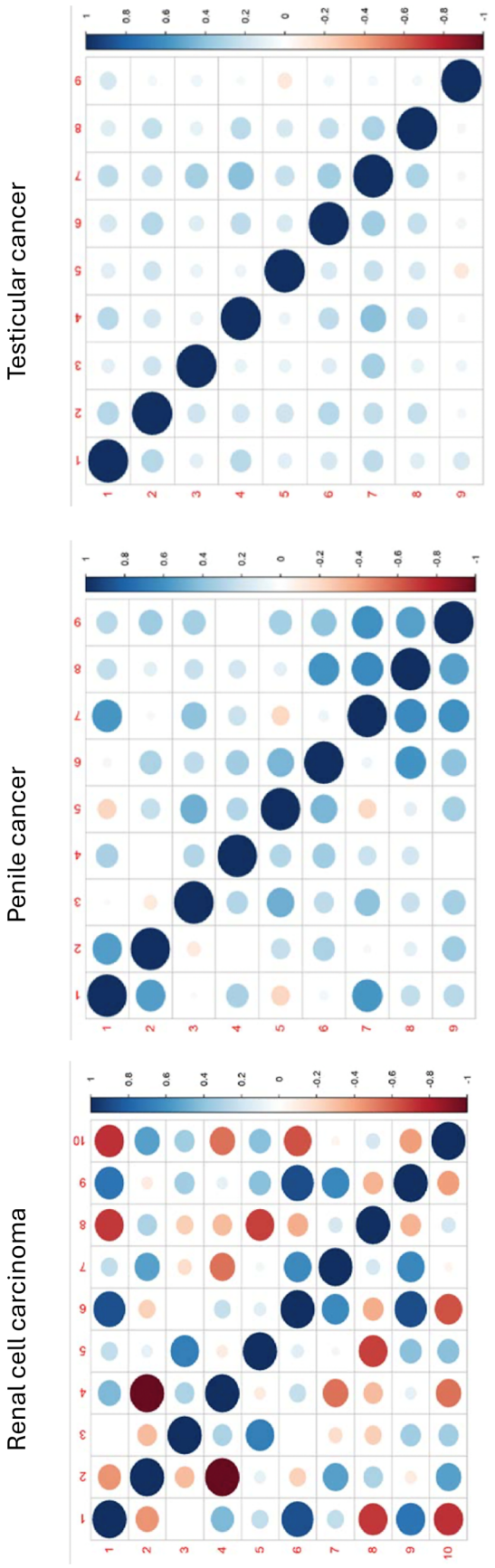
Each heatmap displays pairwise Spearman correlation coefficients between expert ratings of large language model (LLM) responses for guideline-based questions. Each label on the horizontal and vertical axes corresponds to a specific expert who participated in rating the LLM-generated answers. Color intensity indicates the strength of correlation (see scale), with cooler colors denoting higher agreement.

Figure 5. Correlation heatmaps of expert rating agreement in prostate cancer and upper urinary tract urothelial carcinoma.



Each heatmap displays pairwise Spearman correlation coefficients between expert ratings of large language model (LLM) responses for guideline-based questions.
Each label on the horizontal and vertical axes corresponds to a specific expert who participated in rating the LLM-generated answers.
Color intensity indicates the strength of correlation (see scale), with cooler colors denoting higher agreement.

Figure 6. Correlation heatmaps of expert rating agreement in muscle and non-muscle-invasive bladder cancer.



Each heatmap displays pairwise Spearman correlation coefficients between expert ratings of large language model (LLM) responses for guideline-based questions.
Each label on the horizontal and vertical axes corresponds to a specific expert who participated in rating the LLM-generated answers.
Color intensity indicates the strength of correlation (see scale), with cooler colors denoting higher agreement.

Figure 7. Correlation heatmaps of expert rating agreement in renal cell carcinoma, penile cancer, and testicular cancer.

cancers, we identified notable differences in the clinical accuracy and usability of these models. Notably, the simple version of Claude consistently achieved the highest mean scores and the most significant proportion of responses rated as optimal or acceptable by the expert panel, outperforming the other models across most clinical scenarios and tumor types. These findings highlight the promising potential of state-of-the-art LLMs, particularly Claude, to support guideline-based clinical decision-making. However, they also underscore the need for ongoing rigorous validation to ensure the reliability and safety of integrating these tools into routine clinical practice.

Guideline-based questioning is fundamental in urology, as clinical practice guidelines (CPGs) provide a standardized, evidence-based framework for diagnosis, treatment, and patient management. These guidelines, developed by leading organizations such as the European Association of Urology (EAU) and the American Urological Association (AUA), not only synthesize the latest research but also classify the strength of scientific evidence and grade their recommendations accordingly. CPGs support rigorous, evidence-based decision-making and help ensure high-quality, consistent care.[21,22]

CPGs are widely used in urology, with nearly 95% of American urologists reporting use of AUA guidelines, which is linked to improved outcomes and reduced practice variation.[23] The adoption of EAU guidelines is also high across Europe, but real-world adherence remains inconsistent due to factors such as physician experience, patient preferences, institutional resources, local practices, or the complexity of the recommendations.[23–27] The EAU's IMAGINE initiative (IMpact Assessment of Guidelines Implementation and Education) was specifically developed to address this gap by auditing adherence and identifying barriers.[28] For instance, a recent audit revealed inappropriate use of neoadjuvant ADT before prostatectomy, despite clear EAU recommendations against it, illustrating ongoing structural and cognitive implementation gaps.[27]

These findings highlight the potential of LLMs to bridge the gap between clinical guidelines and everyday practice by offering accurate, fast, and accessible responses to complex recommendations. Given the difficulty many clinicians face in navigating and interpreting lengthy guidelines, especially under time constraints, LLMs represent a promising tool to support evidence-based decision-making. This study specifically evaluated whether current LLMs can effectively fulfill that role while staying aligned with the intent of guideline recommendations.

LLMs are advanced artificial intelligence systems built on deep neural network architectures, the transformer, which leverage attention mechanisms to process and generate human-like language. These models are trained on massive datasets comprising diverse and extensive text sources, enabling them to capture complex linguistic patterns, semantic relationships, and contextual meaning. The transformer architecture is central to the success of LLMs, as it enables the efficient handling of long-range dependencies in language, making these models particularly versatile for a wide range of natural language processing tasks.[29,30]

In healthcare, LLMs have been rapidly adopted for applications such as medical education, clinical decision support, and the generation of patient-facing materials. Specifically in urology, LLMs have been evaluated for their ability to generate patient information leaflets, answer clinical questions, and assist with administrative tasks, such as drafting discharge summaries.[31–33] Studies demonstrate that LLMs can generate patient information with accuracy and readability comparable to traditional sources, while also supporting clinicians by streamlining workflows and improving communication with patients. However, these applications also highlight the need for clinician oversight to ensure the accuracy and safety of LLM-generated content, as well as ongoing evaluation of their performance in real-world clinical scenarios.[31–33]

One key limitation of current LLMs is their inability to interpret non-textual content such as figures, tables, and flowcharts, which are common in clinical guidelines. As these models are primarily text-based, important information conveyed in visual formats may be missed or misinterpreted.[34,35] To address this, visual elements must be carefully converted into structured text before being input into the model, thereby preserving clinical context and relationships. A notable methodological strength of our study was the systematic conversion of all figures and tables from the clinical guidelines into structured textual descriptions before inputting them into the LLMs. This process, guided by optimized prompts and followed by expert clinical review, was essential to ensure

that critical information embedded in visual elements was accurately represented and accessible to the models. By taking this approach, we minimized the risk of information loss or misinterpretation that can occur when LLMs are presented with non-textual data, thereby enhancing the reliability and clinical relevance of the model outputs. This strategy not only improved the overall quality of the LLM responses in our study but also highlights the importance of thoughtful data preparation when applying LLMs to complex, multimodal clinical content.

Evaluating LLMs in clinical settings requires a multidimensional approach that includes expert review, quantitative metrics, and bias analysis. One strength of this study is the participation of a large, multidisciplinary group of experts from urology, medical oncology, radiation oncology, nuclear medicine, pathology, radiology, and hospital pharmacy. These professionals were drawn from all major Spanish scientific societies involved in the care of urological cancers. This broad representation enabled the collection of thousands of independent assessments, increasing the reliability of the findings and ensuring that the analysis reflects a comprehensive range of clinical perspectives across the full spectrum of urological cancers. In addition, the study employed Likert scales, a validated and widely accepted method to quantify expert ratings of accuracy, clinical relevance, and safety in LLM-generated responses. This approach, commonly used in clinical AI research, allows subjective judgment to be transformed into quantitative data, ensuring a robust and reproducible evaluation. Although not applied in our study, internationally recognized tools such as DISCERN[36] and PEMAT[37] are frequently cited for assessing the quality and comprehensibility of health information. DISCERN, a validated 16-item instrument rated on a five-point scale, is designed to evaluate reliability, the presentation of treatment options, and overall content quality.[36] PEMAT similarly offers a structured framework for evaluating the understandability and actionability of patient education materials.[37] These tools represent international standards and could be integrated into future research to enhance the evaluation of LLM-generated content in clinical practice.

Our study provides a comprehensive comparative analysis of leading LLMs in their ability to interpret and respond to guideline-based questions across the major urological cancers. Among the six configurations evaluated, the simple version of Claude consistently achieved the highest mean Likert scores, and the largest proportion of responses rated as optimal or acceptable by the expert panel, outperforming Gemini and ChatGPT in most clinical scenarios and tumor types. For example, Claude simple achieved over 45% of responses rated as optimal (Likert 5) and more than 87% rated as optimal or acceptable (Likert 4 or 5) across all guidelines, with particularly strong results in muscle-invasive bladder cancer, penile cancer, and testicular cancer. By contrast, Gemini and ChatGPT, both in their simple and rephrased forms, showed lower rates of optimal scores, though their responses were still concentrated in the higher Likert categories, indicating generally acceptable performance. Notably, rephrased prompts did not consistently improve model performance, and in most cases, the simple versions outperformed their rephrased counterparts. This may be because more detailed or rephrased prompts, while potentially providing more context, can also add unnecessary complexity or verbosity, causing the model to produce longer and sometimes less focused responses, especially when the original prompt already had all the essential information. Since the design and complexity of prompts significantly influence the accuracy and clinical usefulness of LLM responses, careful prompt engineering is essential to optimize model performance and ensure reliable decision support in healthcare.[38]

All models demonstrated variability in accuracy and completeness depending on the complexity of the clinical scenario, the specific cancer type, and the phrasing of the queries. While the best-performing models approached expert-level accuracy in certain domains, they still faced challenges with nuanced questions, negative phrasing, and scenarios that required integration of multidisciplinary evidence, ambiguous wording, or lack of context.[39] These limitations underscore the current need for rigorous expert evaluation and standardized benchmarks to ensure that LLM outputs are clinically reliable and aligned with best practices. Nevertheless, the rapid advances observed, particularly with state-of-the-art models like Claude, suggest that LLMs have significant potential to support clinical decision-making in urology, provided their integration is accompanied by ongoing validation and human oversight.[40,41]

Limitations

The limitations inherent to this study need to be acknowledged and discussed. First, although

guideline figures and tables were converted into text to make them accessible to the LLMs, this process may have introduced bias or omitted visual nuances. Second, the evaluation was based on guideline-derived questions in simulated scenarios, rather than real-time clinical decision-making, which limits generalizability. Third, only three publicly available, low-cost LLMs, Claude, Gemini, and ChatGPT, were assessed; therefore, the results may not apply to other models or future versions. In addition, as all evaluations were conducted in Spanish while the original guidelines were in English, some nuances may have been lost in translation. Furthermore, while the expert panel was multidisciplinary and experienced, there was no external validation and no assessment of clinical impact. These will be addressed in a forthcoming Delphi study to validate the top-performing models in real-world uro-oncology settings. In addition, while the inclusion of a small subset of guideline-unanswerable items provided indirect insight into abstention versus fabrication under guideline-only constraints, this component was not designed or powered as a dedicated hallucination-sensitivity experiment. Lastly, fine-tuning (including continued pretraining, supervised fine-tuning, and preference tuning) is an established strategy for domain adaptation and for reinforcing guideline-conformant behaviors. In this study, we intentionally avoided fine-tuning to benchmark intrinsic model performance under controlled, guideline-only inputs; this reflects study scope rather than a limitation of fine-tuning itself.

Despite these limitations, the study stands out for its systematic design, the involvement of a large multidisciplinary expert panel, and the use of a rigorous and reproducible methodology. This provides a strong foundation for future evaluations and the development of more reliable AI tools to support clinical decision-making in urology.

Conclusion

In conclusion, this study demonstrates that among the six LLM configurations evaluated, the simple version of Claude consistently achieved the highest accuracy and the largest proportion of optimal or acceptable responses across all major urological cancers. Claude rephrased also performed strongly, ranking second in most categories. While all models showed generally acceptable performance, Claude models clearly outperformed Gemini and ChatGPT, especially in complex scenarios such as muscle-invasive bladder cancer, penile cancer, and testicular cancer. However, it is important to point out that the superior performance of Claude may be partially explained by its more recent release compared to the other models, and future evaluations may yield different results as newer versions and alternative models become available. As the field evolves rapidly, continuous benchmarking and re-evaluation will be essential.

These findings underscore the rapid progress of advanced LLMs and their growing potential to support evidence-based decision-making in urology, provided their use is accompanied by ongoing expert validation and human oversight. While LLMs offer an attractive solution for delivering rapid, guideline-based answers, especially for less experienced clinicians who may struggle with complex or lengthy recommendations, these tools must be used responsibly. They are best positioned as decision-support aids that complement, rather than replace, clinical expertise. To ensure safe and effective patient care, their output must always be critically appraised and supervised by qualified professionals.

Declarations

Ethics approval and consent to participate

Not applicable, the study did not involve human or animal subjects.

Consent for publication

Not applicable.

Author contributions

Ángel Borque-Fernando: Conceptualization; Data curation; Formal analysis; Funding acquisition; Investigation; Methodology; Project administration; Resources; Software; Supervision; Validation; Visualization; Writing – original draft; Writing – review & editing.

Denis Navarro: Conceptualization; Data curation; Formal analysis; Funding acquisition; Investigation; Methodology; Project administration; Resources; Software; Supervision; Validation; Visualization; Writing – original draft; Writing – review & editing.

Manuel Doblare: Conceptualization; Data curation; Formal analysis; Funding acquisition; Investigation; Methodology; Project administration; Resources; Software; Supervision; Validation;

Visualization; Writing – original draft; Writing – review & editing.

Luis M. Esteban: Conceptualization; Data curation; Formal analysis; Funding acquisition; Investigation; Methodology; Project administration; Resources; Software; Supervision; Validation; Visualization; Writing – original draft; Writing – review & editing.

Daniel Perez-Fentes: Conceptualization; Data curation; Formal analysis; Funding acquisition; Investigation; Methodology; Project administration; Resources; Software; Supervision; Validation; Visualization; Writing – original draft; Writing – review & editing.

Mario Álvarez-Maestro: Conceptualization; Data curation; Formal analysis; Funding acquisition; Investigation; Methodology; Project administration; Software; Supervision; Validation; Visualization; Writing – original draft; Writing – review & editing.

Rafael A. Medina-López: Conceptualization; Data curation; Formal analysis; Funding acquisition; Investigation; Methodology; Project administration; Resources; Software; Supervision; Validation; Visualization; Writing – original draft; Writing – review & editing.

Oscar Rodríguez Faba: Conceptualization; Data curation; Formal analysis; Funding acquisition; Investigation; Methodology; Project administration; Resources; Software; Supervision; Validation; Visualization; Writing – original draft; Writing – review & editing.

José Rubio-Briones: Conceptualization; Data curation; Formal analysis; Funding acquisition; Investigation; Methodology; Project administration; Resources; Software; Supervision; Validation; Visualization; Writing – original draft; Writing – review & editing.

Sergio Fernández-Pello: Conceptualization; Data curation; Formal analysis; Funding acquisition; Investigation; Methodology; Project administration; Resources; Software; Supervision; Validation; Visualization; Writing – original draft; Writing – review & editing.

Jesús María Fernández-Gómez: Conceptualization; Data curation; Formal analysis; Funding acquisition; Investigation; Methodology; Project administration; Resources; Software; Supervision; Validation; Visualization; Writing – original draft; Writing – review & editing.

Tomás Fernández Aparicio: Conceptualization; Data curation; Formal analysis; Funding acquisition; Investigation; Methodology; Project administration; Resources; Software; Supervision; Validation; Visualization; Writing – original draft; Writing – review & editing.

Félix Guerrero Ramos: Conceptualization; Data curation; Formal analysis; Funding acquisition; Investigation; Methodology; Project administration; Resources; Software; Supervision; Validation; Visualization; Writing – original draft; Writing – review & editing.

Laura Izquierdo: Conceptualization; Data curation; Formal analysis; Funding acquisition; Investigation; Methodology; Project administration; Resources; Software; Supervision; Validation; Visualization; Writing – original draft; Writing – review & editing.

José Luis Álvarez-Ossorio Fernández: Conceptualization; Data curation; Formal analysis; Funding acquisition; Investigation; Methodology; Project administration; Resources; Software; Supervision; Validation; Visualization; Writing – original draft; Writing – review & editing.

Acknowledgements

The authors would like to express their sincere gratitude to the experts who participated in the different boards and contributed their knowledge to the evaluation of the questions. *For prostate cancer*: Teresa Alonso Gordoa, Fernando López Campos, Antonio Rodríguez Fernández, Eugenia García Fernández, Joan Carles Vilanova Busquets, Ángel Borque Fernando, Sara MartÍnez Breijo, EstefanÍa Linares Espinós, María José Ledo Cepero, Juan Gómez Rivas, María Espinosa. *For upper tract urothelial carcinoma*: Javier Martínez Benavides, Richard Mast, Tomás Fernández Aparicio, Marta Trassierra Villa, Alberto Pérez Lanzac, Luis Llanes, and Ana Loizaga. *For muscle-invasive bladder cancer*: María José Juan Fita, Felipe Couñago, Julián Sanz Ortega, Javier Martínez, Jesús María Fernández Gómez, Luis Ladaria Sureda, Felipe Villacampa Auba, Alberto Carrión Puig, Óscar Buisán, and Eloy Vivas. *For non-muscle-invasive bladder cancer*: Javier Puente, Antonio López-Beltran, Javier Martínez, Paula Pelechano Gómez, Félix Guerrero Ramos, José Daniel Subiela Henriquez, José Luis Domínguez Escrig, Toni Vilaseca Cabo, and Ana Plata Bello. *For renal cell carcinoma*: Eliseo Carrasco, Ferrán Algaba, María José Agustín, Carlos Nicolau, Sergio Fernández-Pello, Ignacio Osman, Carmen Mir, Rocío

Barrabino, Bernardo Herrera, Guillermo de Velasco, and Marc Simo Perdigo. *For penile cancer*: Alfonso Gómez de Liaño, Marian Gómez Aparicio, Macarena Rodríguez Fraile, Rafael Luque Barona, José Rubio Briones, Antonio Salinas, Josep Maria Gayà, Antonio Rosino, and Jorge García Rodríguez. *For testicular cancer*: Urbano Anido, Abraham Ocanto, José Lorenzo Muñoz Iglesias, Pilar González-Peramato, Laura Izquierdo, Oscar Gorria Cardesa, Mario Domínguez Esteban, Patricia Ramírez Rodríguez-Bermejo, and Juan Alonso Cabo. The authors also wish to thank Dr. Pablo Rivas for his support with the medical writing on behalf of Content Ed Net. We would also like to express our gratitude to Manuel Espárrago for his technical collaboration on the project's evaluation tools, and to Andrés Mena for his contribution to the design of the user interface and the app for accessing the IA tool.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This project was funded by the Spanish Association of Urology, which also supported the medical writing of the manuscript through Content Ed Net.

Competing interests

The authors declare that there is no conflict of interest.

Availability of data and materials

Not applicable.

ORCID iDs

Ángel Borque Fernando https://orcid.org/0000-0003-0178-4567
Luis M. Esteban https://orcid.org/0000-0002-3007-302X

Supplemental material

Supplemental material for this article is available online.

References

1. GPT-4o. https://openai.com/es-ES/index/hello-gpt-4o/ (2024, accessed 23 June 2025).
2. Google Gemini. https://gemini.google.com (2024, accessed 23 June 2025).
3. Meet Claude. Anthropic. https://www.anthropic.com/claude (2024, accessed 23 June 2025).
4. Du X, Zhou Z, Wang Y, et al. Generative large language models in electronic health records for patient care since 2023: a systematic review. *medRxiv*. Epub ahead of print August 2024. DOI: 10.1101/2024.08.11.24311828.
5. Riedemann L, Labonne M and Gilbert S. The path forward for large language models in medicine is open. *npj Digit Med* 2024; 7(1): 339.
6. EAU Guidelines - Uroweb, https://uroweb.org/guidelines (2025, accessed 19 May 2025).
7. Musheyev D, Pan A, Loeb S, et al. How well do artificial intelligence chatbots respond to the top search queries about urological malignancies? *Eur Urol* 2024; 85(1): 13–16.
8. Wang L, Wan Z, Ni C, et al. Applications and concerns of ChatGPT and other conversational large language models in health care: systematic review. *J Med Internet Res* 2024; 26: e22769.
9. Nori H, Daswani M, Kelly C, et al. Sequential diagnosis with language models. *arXiv*. Preprint posted online 2025. doi:10.48550/ARXIV.2506.22405
10. Rewthamrongsris P, Burapacheep J, Trachoo V, et al. Accuracy of large language models for infective endocarditis prophylaxis in dental procedures. *Int Dental J* 2025; 75(1): 206–212.
11. EAU Guidelines on Prostate Cancer - Uroweb, https://uroweb.org/guidelines/prostate-cancer (2025, accessed 19 May 2025).
12. EAU Guidelines on Upper Urinary Tract Urothelial Cell Carcinoma - Uroweb, https://uroweb.org/guidelines/upper-urinary-tract-urothelial-cell-carcinoma (2025, accessed 19 May 2025).
13. EAU Guidelines on Muscle-invasive and Metastatic Bladder Cancer - Uroweb, https://uroweb.org/guidelines/muscle-invasive-and-metastatic-bladder-cancer (2025, accessed 19 May 2025).
14. EAU Guidelines on Non-muscle-invasive Bladder Cancer - Uroweb, https://uroweb.org/guidelines/non-muscle-invasive-bladder-cancer (2025, accessed 19 May 2025).
15. EAU Guidelines on Testicular Cancer – Uroweb, https://uroweb.org/guidelines/testicular-cancer (2025, accessed 19 May 2025).
16. EAU Guidelines on Renal Cell Carcinoma - Uroweb, https://uroweb.org/guidelines/renal-cell-carcinoma (2025, accessed 19 May 2025).
17. EAU Guidelines on Penile Cancer - Uroweb, https://uroweb.org/guidelines/penile-cancer (2023, accessed 19 May 2025).
18. Meskó B. Prompt engineering as an important emerging skill for medical professionals: tutorial. *J Med Internet Res* 2023; 25: e50638.

19. Patil R, Heston TF and Bhuse V. Prompt engineering in healthcare. *Electronics* 2024; 13(15): 2961.
20. Sarangi PK, Datta S, Swarup MS, et al. Radiologic decision-making for imaging in pulmonary embolism: accuracy and reliability of large language models—Bing, Claude, ChatGPT, and Perplexity. *Indian J Radiol Imaging* 2024; 34(04): 653–660.
21. Methodology & Polices - Uroweb, https://uroweb.org/eau-guidelines/methodology-policies (2025, accessed 19 May 2025).
22. Standard Operating Procedures Overview - American Urological Association, https://www.auanet.org/guidelines-and-quality/guidelines/standard-operating-procedures-overview (2025, accessed 19 May 2025).
23. Breyer BN, Fang R, Meeks W, et al. Use of the American Urological Association Clinical Practice Guidelines: data from the AUA census. *Urol Pract* 2017; 4(6): 462–467.
24. Bada M, Berardinelli F, Nyirády P, et al. Adherence to the EAU guidelines on penile cancer treatment: European, multicentre, retrospective study. *J Cancer Res Clin Oncol* 2019; 145(4): 921–926.
25. Cacciamani G, Artibani W, Briganti A, et al. Adherence to the European Association of Urology Guidelines: A National Survey among Italian urologists. *Urol Int* 2018; 100(2): 139–145.
26. Hendricksen K, Aziz A, Bes P, et al. Discrepancy between european association of urology guidelines and daily practice in the management of non–muscle-invasive bladder cancer: results of a European Survey. *Eur Urol Focus* 2019; 5(4): 681–688.
27. MacLennan S, Azevedo N, Duncan E, et al. Mapping European Association of Urology Guideline Practice Across Europe: an audit of androgen deprivation therapy use before prostate cancer surgery in 6598 cases in 187 hospitals across 31 European Countries. *Eur Urol* 2023; 83(5): 393–401.
28. Cornford P, Smith EJ, MacLennan S, et al. IMAGINE-IMpact assessment of guidelines implementation and education: the next frontier for harmonising urological practice across europe by improving adherence to guidelines. *Eur Urol* 2021; 79(2): 173–176.
29. Large Language Models are Transformers in Artificial Intelligence, Industry, Education, and Society, 2024.
30. Raza M, Jahangir Z, Riaz MB, et al. Industrial applications of large language models. *Sci Rep* 2025; 15(1): 13755.
31. Mak G, Siriwardena C, Haxhimolla H, et al. Utility of ChatGPT and large language models in enhancing patient understanding of urological conditions. *SIUJ* 2024; 5(6): 843–851.
32. Pompili D, Richa Y, Collins P, et al. Using artificial intelligence to generate medical literature for urology patients: a comparison of three different large language models. *World J Urol* 2024; 42(1): 455.
33. Zhang L, Zhao Q, Zhang D, et al. Application of large language models in healthcare: a bibliometric analysis. *Digit Health* 2025; 11: 20552076251324444.
34. Pang C, Cao Y, Yang C, et al. Uncovering limitations of large language models in information seeking from tables. In: *Findings of the association for computational linguistics ACL* 2024, 2024, pp. 1388–1409. Association for Computational Linguistics.
35. Lu W, Zhang J, Fan J, et al. Large language model for table processing: a survey. *Front Comput Sci* 2025; 19(2): 192350.
36. 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(2): 105–111.
37. Furukawa E, Okuhara T, Liu M, et al. Evaluating online and offline health information with the patient education materials assessment tool: protocol for a systematic review. *JMIR Res Protoc* 2025; 14: e63489.
38. Sarangi PK and Mondal H. Response generated by large language models depends on the structure of the prompt. *Indian J Radiol Imaging* 2024; 34(3): 574–575.
39. Wang Y, Zhao Y and Petzold L. Are large language models ready for healthcare? a comparative study on clinical language understanding. In: *Proceedings of the 8th machine learning for healthcare conference*, PMLR 219, pp. 804–823, 2023.
40. Harskamp RE and De Clercq L. Performance of ChatGPT as an AI-assisted decision support tool in medicine: a proof-of-concept study for interpreting symptoms and management of common cardiac conditions (AMSTELHEART-2). *Acta Cardiol* 2024; 79(3): 358–366.
41. Li KP, Wang L, Wan S, et al. Enhanced artificial intelligence in bladder cancer management: a comparative analysis and optimization study of multiple large language models. *J Endourol* 2025; 39(5): 494–499.