Research Article

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



Journal of **Comparative Effectiveness Research**

Evaluation of large language models for guideline-based clinical questions in muscle-invasive and non-muscle-invasive bladder cancer and upper tract urothelial carcinoma: a multidisciplinary Delphi study

Félix Guerrero-Ramos[‡,1], Jesús María Fernández Gómez[‡,2], Tomás Fernández Aparicio[‡,3], Antonio López-Beltrán[4], Óscar Buisán Rueda[5], Luis Llanes González[6], Javier Puente Vázquez[7], Alberto Carrión[8], Ana Loizaga Iriarte[9], José Daniel Subiela Henríquez[10], Felipe Couñago[11], Javier Martínez Benavides[12], Paula Pelechano Gómez[13], María José Juan Fita[14], Richard Mast Vilaseca[15], Antoni Vilaseca[16], Luis Ladaria Sureda[17], Alberto Pérez-Lanzac de Lorca[18], José Luis Domínguez Escrig[19], Julián Sanz Ortega[20], Elena Sevillano Fernández[21], Ana Cristina Plata Bello[22], Felipe Villacampa Auba[23], Marta Trassierra Villa[24], Eloy Vivas Vargas[25], Isabel Trias Puigsureda[26] & Ángel Borque-Fernando*[,27]

[1]Urology Department, Hospital Universitario 12 de Octubre, Madrid, Spain
[2]Urology Department, Hospital Universitario Central de Asturias, Universidad de Oviedo, Oviedo, Spain
[3]Urology Department, Hospital General Universitario Morales Meseguer, Murcia, Spain
[4]Department of Morphological Sciences, Anatomical Pathology Unit, Universidad de Córdoba, Córdoba, Spain
[5]Urology Department, Hospital Universitari de Bellvitge, L'Hospitalet de Llobregat, Barcelona, Spain
[6]Urology Department, Hospital Universitario de Getafe, Getafe, Madrid, Spain
[7]Medical Oncology Department, Hospital Clínico San Carlos, Madrid, Spain
[8]Urology Department, Hospital Universitari Vall d'Hebron, Barcelona, Spain
[9]Urology Department, Hospital Universitario Basurto, Bilbao, Spain
[10]Urology Department, Hospital Universitario Ramón y Cajal, Madrid, Spain
[11]Department of Medicine, Faculty of Medicine, Health & Sports, Universidad Europea de Madrid, Madrid, Spain; Radiation Oncology Department, GenesisCare Madrid, Hospital Universitario San Francisco de Asís & Hospital Universitario Vithas Madrid La Milagrosa, Madrid, Spain
[12]Pharmacy Department, Institut Català d'Oncologia, Barcelona, Spain
[13]Radiology Department, Fundación Instituto Valenciano de Oncología, Valencia, Spain
[14]Medical Oncology Department, Fundación Instituto Valenciano de Oncología, Valencia, Spain
[15]Radiology Department, Hospital Universitari Vall d'Hebron, Barcelona, Spain
[16]Urology Department, Hospital Clínic de Barcelona, Barcelona, Spain
[17]Urology Department, Hospital Universitari Son Espases, Palma de Mallorca, Spain
[18]Department of Medicine, Faculty of Medicine, Health & Sports, Universidad Europea de Madrid. Urology Department, Hospital Universitario Vithas Madrid Arturo Soria and APL Urología. Madrid, Spain
[19]Urology Department, Fundación Instituto Valenciano de Oncología, Valencia, Spain
[20]Pathology Department, Clínica Universidad de Navarra, Madrid, Spain
[21]Centro Integral Oncológico Clara Campal (HM CIOCC), Hospital Universitario HM Sanchinarro, Madrid, Spain
[22]Urology Department, Hospital Universitario de Canarias, La Laguna, Santa Cruz de Tenerife, Spain
[23]Urology Department, Clínica Universidad de Navarra, Madrid, Spain
[24]Urology Department, Hospital Clínico Universitario de Valencia, Valencia, Spain
[25]Radiology Department, Hospital Universitario Costa del Sol, Marbella, Málaga, Spain
[26]Pathology Department, Hospital Clínic de Barcelona, Barcelona, Spain
[27]Urology Department, Hospital Universitario Miguel Servet, IIS Aragón (GIIS-071), Zaragoza, Spain
*Author for correspondence: dr.borque@gmail.com
‡Authors contributed equally as co-first authors

Aim: To assess expert panel agreement with responses from three large language models, Claude, Gemini and ChatGPT, to questions based on guidelines for non-muscle-invasive bladder cancer (NMIBC), muscle-invasive bladder cancer (MIBC) and upper tract urothelial carcinoma (UTUC). **Materials & methods:** The 2025 European Association of Urology (EAU) guidelines were provided to Claude 3.5 Haiku, Gemini 2.5 Flash and GPT-4.1-mini. Figures and tables were converted into plain language text. Each model answered

25 questions per entity. Panels of 7–10 experts rated responses on a 9-point scale using a modified Delphi process with two rounds. Outcomes were consensus in agreement and median expert panel score ≥7. **Results:** For Claude, Gemini and ChatGPT, respectively, consensus in agreement was achieved for 88%, 100% and 96% of NMIBC responses ($p = 0.097$); 92%, 100% and 96% of MIBC responses ($p = 0.223$); and 56%, 100% and 96% of UTUC responses ($p < 0.001$). For UTUC, significantly fewer Claude responses achieved consensus in agreement than Gemini or ChatGPT responses. Across all questions, consensus in agreement was achieved for 59/75 Claude responses (78.7%), 75/75 Gemini responses (100%), and 72/75 ChatGPT responses (96%) ($p < 0.001$). Median scores ≥7 were achieved for 88%, 100% and 100% of NMIBC responses; 96%, 100% and 100% of MIBC responses; and 68%, 100% and 96% of UTUC responses. In UTUC, Claude had significantly fewer such responses than Gemini, but the difference from ChatGPT was not statistically significant. Overall, median scores ≥7 were achieved for 63/75 Claude responses (84%), 75/75 Gemini responses (100%) and 74/75 ChatGPT responses (98.7%) ($p < 0.001$). **Conclusion:** Under the study conditions, large language model responses frequently achieved expert panel agreement, particularly for Gemini and ChatGPT, supporting the potential of these models for guideline interpretation. Further validation is required to establish clinical utility and safety.

Plain language summary: Artificial intelligence tools for interpreting guidelines on bladder & upper urinary tract cancers

What is this article about? Clinical guidelines help healthcare professionals make decisions about bladder and upper urinary tract cancers, but they can be long and complex. Artificial intelligence tools such as Claude, Gemini and ChatGPT may help interpret guideline information. However, their answers need expert review before clinical use.

What methodology was used in this study? Specialists created 25 questions for three clinical areas: non-muscle-invasive bladder cancer, muscle-invasive bladder cancer and upper tract urothelial carcinoma. The 2025 European Association of Urology guidelines were provided to the three tools, with figures and tables converted into clear text. Each tool answered 75 questions. Panels of 7–10 specialists reviewed the responses in two rounds.

What were the results? Across all 75 questions, experts agreed with 59 Claude responses, 75 Gemini responses and 72 ChatGPT responses. The largest differences occurred in upper tract urothelial carcinoma, where experts agreed with 14 of 25 Claude responses, 25 of 25 Gemini responses and 24 of 25 ChatGPT responses. Expert comments highlighted limitations in the completeness and focus of some Claude responses.

Why are these results important? These tools can often interpret supplied guideline information in a way that agrees with specialist opinion. However, the study did not assess their safety or usefulness in real clinical practice. Their responses should therefore be checked by healthcare professionals, particularly in complex situations.

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

Keywords: artificial intelligence • consensus • large language models • upper urinary tract tumor • urinary bladder neoplasms • urothelial cancer

Urothelial carcinomas (UC) comprise a heterogeneous group of malignancies arising from the urothelium of the urinary tract, including bladder cancer (BC), both non-muscle-invasive bladder cancer (NMIBC) and muscle-invasive bladder cancer (MIBC), as well as upper urinary tract urothelial carcinoma (UTUC) [1,2]. BC represents 90–95% of UCs, with approximately 75% of cases classified as NMIBC and the remaining 25% as MIBC [3,4]. UTUC accounts for 5–10% of all UC cases and presents distinct clinical and epidemiological characteristics [5].

BC ranks as the ninth most common cancer worldwide and the thirteenth leading cause of cancer death [6]. Each year, about 614,000 new cases and 220,000 deaths occur globally [6]. The disease predominantly affects men, with an incidence rate of 9.5 per 100,000 men worldwide and a lower rate of 2.4 per 100,000 women [4]. Europe experiences the highest burden, contributing over a third of cases and deaths [6]. However, BC incidence and mortality rates vary across countries due to differences in risk factors, detection and diagnostic practices and availability of treatments [4]. These figures underscore BC's significant impact on public health and the critical need for effective prevention, early detection and management strategies. Conversely, UTUC is a less common but

clinically significant malignancy, with an estimated annual incidence of around 2–3 cases per 100,000 inhabitants in Western countries [5,7]. It primarily affects individuals aged 70–90 years and occurs twice as frequently in men [5]. Unlike BC, about two-thirds of UTUC cases are muscle-invasive at diagnosis, reflecting its often aggressive behavior [5]. Up to 9% of patients present with metastatic disease, and bladder recurrence after UTUC treatment occurs in nearly 30% of cases [5]. These data highlight the distinct challenges in diagnosing and managing UTUC, emphasizing the importance of a proactive approach and close follow-up.

Clinical practice guidelines, such as those developed by the European Association of Urology (EAU), are essential for standardizing the management of UC, including NMIBC [3], MIBC [4] and UTUC [5]. These documents synthesize current evidence and expert consensus to optimize patient care. However, the complexity of guideline recommendations and the heterogeneity of clinical scenarios can hinder consistent implementation in routine practice. Moreover, the relative rarity of UTUC and the distinct biological behavior of NMIBC versus MIBC pose additional challenges to accurate clinical decision-making. Recently, large language models (LLMs) have emerged as promising tools to facilitate access to up-to-date, guideline-based recommendations and support clinical decision-making [8–13]. Our prior work indicated that using simple, unmodified questions increases the likelihood of generating guideline-consistent responses from LLMs in uro-oncological contexts [14]. While results showed comparable performance across leading models, including Claude, Gemini and ChatGPT, a need remains for rigorous evaluation of these AI systems' consistency, reliability and alignment with expert consensus across diverse urothelial tumor types and clinical scenarios.

This study aimed to assess expert panel agreement with responses generated by Claude, Gemini and ChatGPT to clinical questions based on the EAU guidelines for NMIBC, MIBC and UTUC. Using a modified Delphi methodology, multidisciplinary expert panels independently rated the degree of agreement with each LLM response. This article focuses on NMIBC, MIBC and UTUC, complementing companion papers addressing other major urological malignancies such as penile, testicular, prostate and renal cancers.

Materials & methods

Study design

This multidisciplinary study focused on NMIBC, MIBC and UTUC. We employed a two-round Delphi methodology to assess the degree of 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 for these uro-oncological malignancies, coupled with their favorable 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 boards and three expert panels: the Scientific Board provided strategic oversight and ensured broad representation across the major urological cancer subtypes; the Technical Board provided methodological and digital expertise and three Specialized Tumor Boards, one for each UC clinical entity, included multidisciplinary experts: 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 Anatomical Pathology [SEAP], Spanish Society of Hospital Pharmacy [SEFH] and Spanish Society of Medical Radiology [SERAM]). Panelists were nominated by the participating scientific societies based on their recognized expertise and received no compensation. Individual years of experience and institutional affiliations were not collected.

Adaptation of clinical guidelines to plain language

All figures and tables from the EAU clinical guidelines for NMIBC [3], MIBC [4] and UTUC [5] were systematically transformed into clear, plain-language textual summaries to facilitate interpretation by the LLMs. The guidelines

were arranged into thematic sections and reviewed by independent members of the specialized UC tumor boards. Any discrepancies were resolved through consensus among the guideline coordinators, with final approval requiring unanimous agreement.

Questionnaire development

The three Specialized Tumor Boards created 25 structured questions for each clinical entity, resulting in 75 questions in total, based on the original EAU guideline content. These questions covered key clinical topics such as diagnosis, risk assessment, indications for diagnostic tests and treatment options. Members of the specialized tumor boards contributed to question development and subsequently evaluated the LLM responses. The MIBC section followed the full scope of the 2025 EAU Guidelines on Muscle-invasive and Metastatic Bladder Cancer and therefore included three questions addressing metastatic disease (questions 23, 24 and 25).

LLM selection & response generation

Three widely accessible LLMs were tested: Claude [15] (Anthropic; version claude-3-5-haiku-20241022, launched 22 October 2024), Gemini [16] (Google; version gemini-2.5-flash-preview-05-20, launched 20 May 2025) and ChatGPT [17] (OpenAI; version gpt-4.1-mini-2025-04-14, launched 14 April 2025). Each model was evaluated using a simple prompt approach. The original 25 structured questions were submitted without modification in Spanish, independently and one at a time, and each model was evaluated separately rather than within a single continuous conversation. One response was generated per question and model. The original EAU guideline text was provided in English without translation. Identical 2025 EAU guideline materials, including the same plain-language textual summaries of figures and tables, were provided to all three models. No input length limitations were encountered, and no truncation was required. Thus, the study evaluated responses generated from supplied guideline content rather than independent clinical decision-making.

Delphi validation process & consensus criteria

Expert panelists independently assessed each LLM response using a 9-point Likert scale (1 = strong disagreement, 9 = strong agreement) to evaluate alignment with EAU guideline content. Panelists were blinded to the identity of the LLM that generated each response.

The Delphi process consisted of two rounds. In the first round, all responses were scored. Questions without consensus after the initial round were reevaluated in a second round, which included distributing anonymized, aggregated feedback summarizing group ratings. Consensus was defined based on three criteria: the median score falling within an agreement (7–9) or disagreement (1–3) range; fewer than a third of panelists having ratings outside these ranges and an interquartile range (IQR) less than 4. Qualitative comments were recorded to provide context for the quantitative results.

The LLM response text remained unchanged between rounds; only panelist ratings could be revised after reviewing the anonymized, aggregated feedback. Panelists could abstain from rating responses outside their area of expertise. Consensus percentages and median scores were calculated using only the ratings submitted for each response, excluding abstentions.

Statistical analysis

Two binary outcomes were considered for model comparison: achievement of consensus in agreement versus lack of consensus in agreement, which included both no consensus and consensus in disagreement, and a median score ≥ 7 versus <7. Because all three LLMs were evaluated using the same clinical questions (25 per clinical entity and 75 in the combined analysis), analyses accounted for the paired structure of the data. Overall differences among the three models were assessed using Cochran's Q test, while pairwise model comparisons were performed using exact McNemar tests. Bonferroni correction was applied to the three pairwise comparisons within each outcome, resulting in a significance threshold of $p < 0.0167$. Overall comparisons among NMIBC, MIBC and UTUC, which comprised independent sets of questions, were performed using Fisher's exact test. Overall comparisons were considered statistically significant at $p < 0.05$. Proportions are reported with 95% CIs calculated using the Wilson method.

The full set of responses generated by each LLM to all 75 questions is available for consultation at https://www.contentednet.com/proyecto-urooncoguias-aeu/.

Results

Non-muscle-invasive bladder cancer

The number of panelists rating each LLM response ranged from seven to nine. All panelists participated in both Delphi rounds. The questions, details of the consensus results, and the main summarized panelist comments are shown in Supplementary Table 1. The number of panelists, median score, and interquartile range for each question and LLM are reported in Supplementary Table 2. Median scores assigned by panelists for each response and LLM are shown in Figure 1. The proportion of panelists scoring in each category (1–3 = disagreement, 4–6 = no consensus, 7–9 = agreement) per clinical question and LLM is shown in Supplementary Figure 1.

For Claude, consensus in agreement was reached in the first round for 18 responses (72%), in the second round for 4 responses that had not achieved prior consensus (16%), while consensus was not reached for 2 responses (8%). Consensus in disagreement was reached in the first round for one response (4%). Overall, consensus in agreement was reached for 22 out of 25 items (88%, 95% CI: 70.0–95.8%). For Gemini, consensus in agreement was reached in the first round for 23 responses (92%), in the second round for 2 responses (8%). No responses remained without consensus. Overall, consensus in agreement was achieved for all 25 responses (100%, 95% CI: 86.7–100.0%). For ChatGPT, consensus in agreement was reached in the first round for 18 responses (72%) and in the second round for 6 responses (24%), while consensus was not reached for 1 response (4%). Overall, consensus in agreement was reached on 24 out of 25 items (96%, 95% CI: 80.5–99.3%).

No statistically significant differences were found in the proportion of items reaching consensus in agreement among the three LLMs, either overall ($p = 0.097$) or in pairwise comparisons (Figure 2A). Median scores ≥ 7 were achieved for 22/25 Claude responses (88%, 95% CI: 70.0%–95.8%) and all 25 Gemini and ChatGPT responses (100%, 95% CI: 86.7–100.0%). The overall comparison was statistically significant ($p = 0.0498$), but none of the pairwise comparisons reached the Bonferroni-corrected significance threshold (Figure 2B).

Muscle-invasive bladder cancer

The number of panelists rating each LLM response ranged from seven to ten. All panelists participated in both Delphi rounds. The questions, details of the consensus results, and the main summarized panelist comments are shown in Supplementary Table 3. The number of panelists, median score and interquartile range for each question and LLM are reported in Supplementary Table 4. Median scores assigned by panelists for each response and LLM are shown in Figure 3. The proportion of panelists scoring in each category (1–3 = disagreement, 4–6 = no consensus, 7–9 = agreement) per response and LLM is shown in Supplementary Figure 2.

For Claude, consensus in agreement was reached in the first round for 22 responses (88%), in the second round for one response that had not achieved prior consensus (4%), while consensus was not reached for 2 responses (8%). Overall, consensus in agreement was reached for 23 out of 25 items (92%, 95% CI: 75.0–97.8%). For Gemini, consensus in agreement was reached in the first round for 23 responses (92%), in the second round for two responses that had not achieved prior consensus (8%). No responses remained without consensus. Overall, consensus in agreement was achieved for all 25 responses (100%, 95% CI: 86.7–100.0%). For ChatGPT, consensus in agreement was reached in the first round for 23 responses (92%), in the second round for one response that had not achieved prior consensus (4%), and consensus was not reached for 1 response (4%). Overall, consensus in agreement was reached for 24 out of 25 items (96%, 95% CI: 80.5–99.3%). No consensus in disagreement was reached on any response from any of the three models. No statistically significant differences were observed in the proportion of items reaching consensus in agreement among the three LLMs, either overall ($p = 0.223$) or in pairwise comparisons (Figure 4A). Median scores ≥ 7 were achieved for 24/25 Claude responses (96%, 95% CI: 80.5–99.3%) and all 25 Gemini and ChatGPT responses (100%, 95% CI: 86.7–100.0%). No statistically significant differences were identified either overall ($p = 0.368$) or in pairwise comparisons (Figure 4B).

The responses for which no consensus was reached concerned the use of magnetic resonance imaging and neoadjuvant therapy. Comments indicated that the responses did not fully align with current evidence (Supplementary Table 3).

Upper tract urothelial carcinoma

The number of panelists rating each LLM response ranged from five to seven. All panelists participated in both Delphi rounds. The questions, details of the consensus results and the main summarized panelist comments are shown in Supplementary Table 5. The number of panelists, median score and interquartile range for each question and LLM are reported in Supplementary Table 6. Median scores assigned by panelists for each response and LLM are

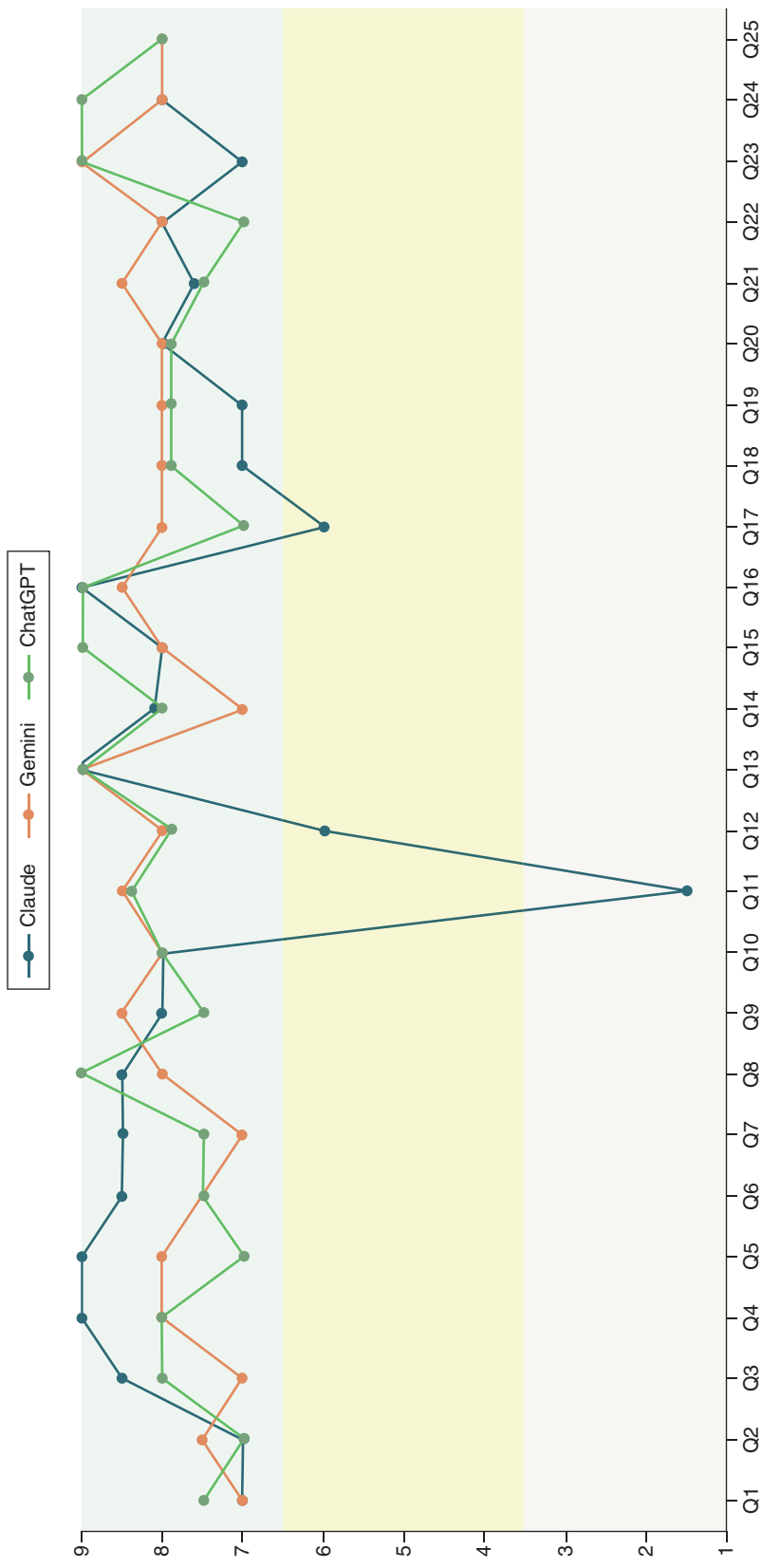


Figure 1. Median scores assigned by panelists for each response and large language model in non-muscle-invasive bladder cancer. Median scores on a 9-point scale are shown for each clinical question and large language model (Claude, Gemini and ChatGPT). Score ranges were defined as: 1–3 = disagreement, 4–6 = no consensus, 7–9 = agreement. Clinical questions are discrete and not sequentially ordered; connecting lines are used only to facilitate visual comparison among models and do not represent trends across questions. The number of panelists, median score and interquartile range for each question and model are reported in Supplementary Table 2.

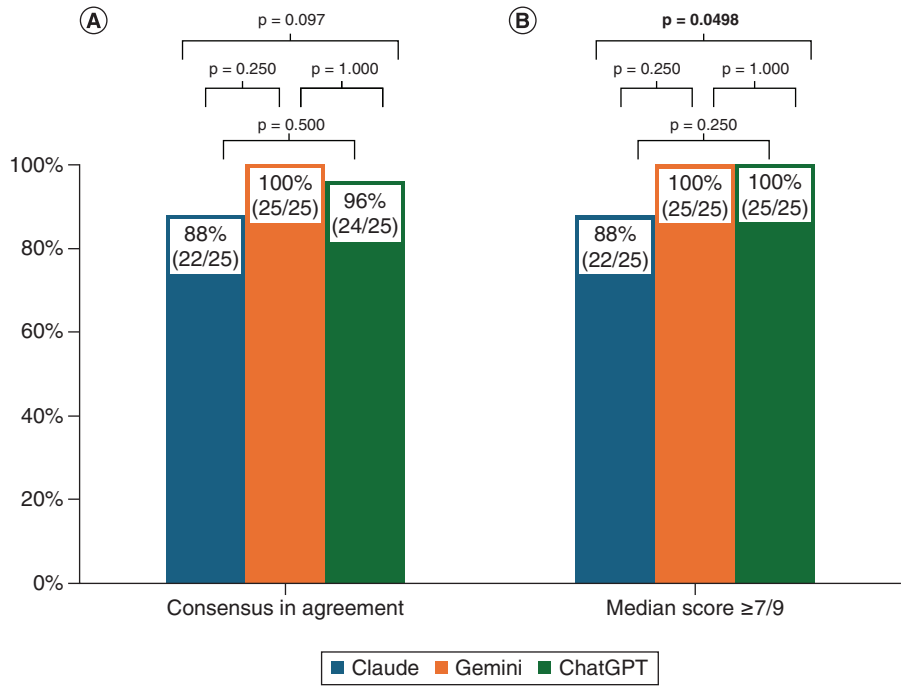


Figure 2. Comparison of expert panel agreement across large language models in non-muscle-invasive bladder cancer. (A) Proportion of consensus in agreement and **(B)** high-scoring items (median ≥7/9) across 25 responses for each large language model in non-muscle-invasive bladder cancer. Pairwise comparisons were considered statistically significant at $p < 0.0167$ after Bonferroni correction.

shown in Figure 5. The proportion of panelists scoring in each category (1–3 = disagreement, 4–6 = no consensus, 7–9 = agreement) per clinical question and LLM is shown in Supplementary Figure 3.

For Claude, consensus in agreement was reached in the first round for 12 responses (48%), in the second round for two responses that had not achieved prior consensus (8%), while consensus was not reached for 11 responses (44%). Overall, consensus in agreement was reached for 14 out of 25 items (56%, 95% CI: 37.1–73.3%). For Gemini, consensus in agreement was achieved in the first round for all 25 responses (100%, 95% CI: 86.7–100.0%). For ChatGPT, consensus in agreement was reached in the first round for 23 responses (92%), in the second round for one response that had not achieved prior consensus (4%), and consensus was not reached for one response (4%). Overall, consensus in agreement was reached for 24 out of 25 items (96%, 95% CI: 80.5–99.3%). No consensus in disagreement was reached on any response from any of the three models.

There were statistically significant differences among the three LLMs in both consensus in agreement ($p < 0.001$) and the proportion of high-scoring responses (median score ≥7/9) ($p = 0.0018$) for UTUC (Figure 6). Consensus in agreement was achieved for 100% of Gemini responses (25/25), 96% of ChatGPT responses (24/25) and 56% of Claude responses (14/25). Significantly fewer Claude responses achieved consensus in agreement than Gemini responses ($p < 0.001$) or ChatGPT responses ($p = 0.0020$), while no significant difference was identified between Gemini and ChatGPT (Figure 6A). Median scores ≥7 were achieved for 17/25 Claude responses (68%, 95% CI: 48.4–82.8%), 25/25 Gemini responses (100%, 95% CI: 86.7–100.0%), and 24/25 ChatGPT responses (96%, 95% CI: 80.5–99.3%). Claude had significantly fewer responses with median scores ≥7 than Gemini ($p = 0.0078$), but the difference between Claude and ChatGPT did not reach the corrected significance threshold ($p = 0.0391$). Gemini and ChatGPT did not differ significantly (Figure 6B).

The clinical management items related to advanced UTUC, in which Claude failed to achieve consensus in agreement, primarily involved complex therapeutic decisions and surgical considerations. These included the use of intravesical chemotherapy after radical nephroureterectomy (item 2), intraluminal chemotherapy post-ureteroscopy without biopsy (item 5) and the extent of lymphadenectomy in ureteral tumors (items 9 and 19). Complex case-based scenarios (items 7, 18 and 20) and specific patient management in the context of syndromes such as Lynch syndrome (item 13) were also challenging. Expert panelists highlighted inaccuracies or incomplete explanations as

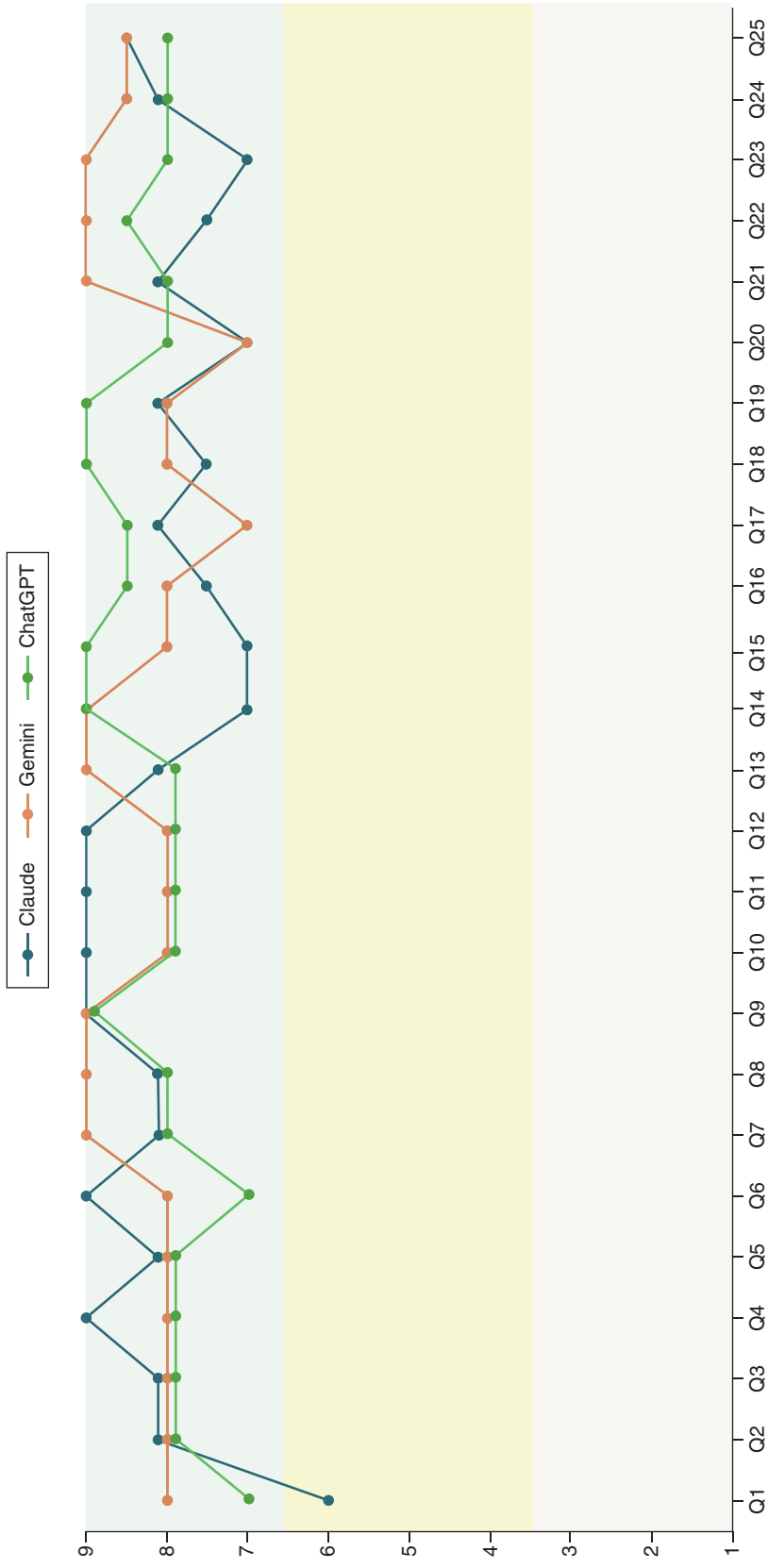


Figure 3. Median scores assigned by panelists for each response and large language model in muscle-invasive bladder cancer. Median scores on a 9-point scale are shown for each clinical question and large language model (Claude, Gemini and ChatGPT). Score ranges were defined as: 1–3 = disagreement, 4–6 = no consensus, 7–9 = agreement. Clinical questions are discrete and not sequentially ordered; connecting lines are used only to facilitate visual comparison among models and do not represent trends across questions. The number of panelists, median score and interquartile range for each question and model are reported in Supplementary Table 4.

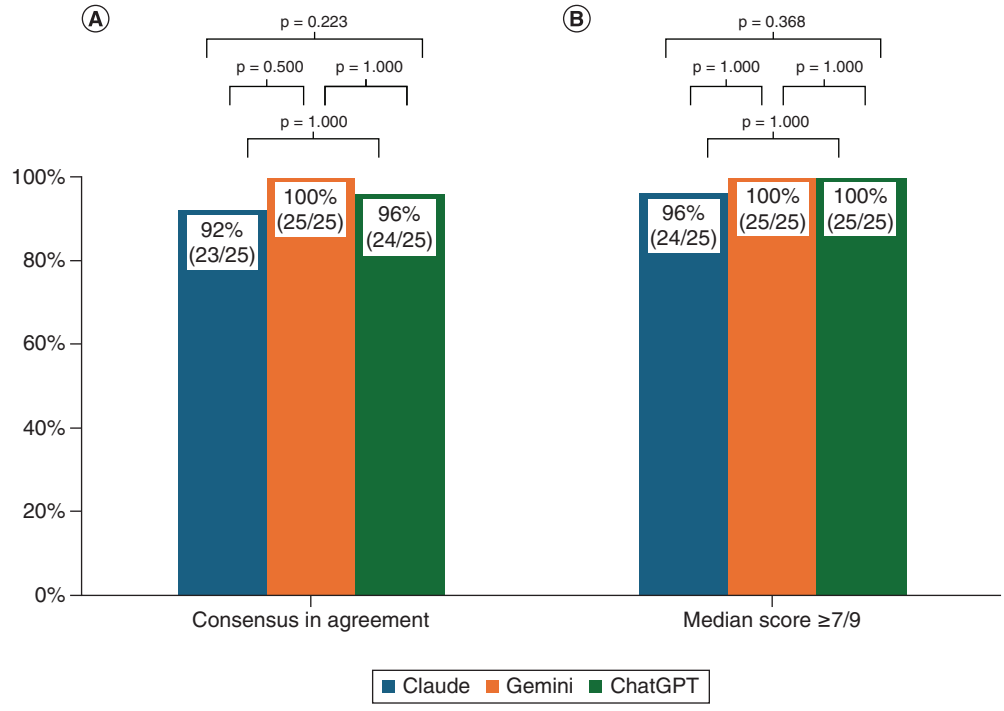


Figure 4. Comparison of expert panel agreement across large language models in muscle-invasive bladder cancer. **(A)** Proportion of consensus in agreement and **(B)** high-scoring items (median ≥7/9) across 25 responses for each large language model in muscle-invasive bladder cancer. Pairwise comparisons were considered statistically significant at $p < 0.0167$ after Bonferroni correction.

the main reasons for discordance. Conversely, panelists described Claude's responses as clear or concise for items 1, 6, 10, 14, 16 and 17, although the response to item 1 was also noted to include incorrect alternatives (Supplementary Table 5).

Combined results for NMIBC, MIBC & UTUC

Across all 75 questions, significantly fewer Claude responses achieved consensus in agreement than Gemini or ChatGPT responses. For this outcome, consensus in agreement was achieved for 75 out of 75 Gemini responses (100%, 95% CI: 95.1–100.0%), 72 out of 75 ChatGPT responses (96%, 95% CI: 88.9–98.6%) and 59 out of 75 Claude responses (78.7%, 95% CI: 68.1–86.4%). The overall comparison was statistically significant ($p < 0.001$). The number of responses lacking consensus in agreement was 16 for Claude, compared with 0 for Gemini and 3 for ChatGPT. Pairwise analyses showed significant differences between Claude and Gemini ($p < 0.001$), as well as between Claude and ChatGPT ($p < 0.001$). However, there was no statistically significant difference between Gemini and ChatGPT ($p = 0.250$).

Across all 75 questions, significantly fewer Claude responses had median scores ≥7 than Gemini or ChatGPT responses. Median scores ≥7 were recorded for all 75 Gemini responses (100%, 95% CI: 95.1–100.0%), 74 out of 75 ChatGPT responses (98.7%, 95% CI: 92.8–99.8%) and 63 out of 75 Claude responses (84%, 95% CI: 74.1–90.6%). The overall comparison was statistically significant ($p < 0.001$). Pairwise analyses showed significant differences between Claude and Gemini ($p < 0.001$), as well as between Claude and ChatGPT ($p = 0.0034$). However, there was no statistically significant difference between Gemini and ChatGPT ($p = 1$).

When the results of the three LLMs were pooled within each clinical entity, consensus in agreement was achieved for 94.7% of NMIBC responses (71/75), 96.0% of MIBC responses (72/75) and 84.0% of UTUC responses (63/75), with a significant overall difference ($p = 0.0243$). Median scores ≥7 were achieved for 96.0% of NMIBC responses (72/75), 98.7% of MIBC responses (74/75) and 88.0% of UTUC responses (66/75), also with a significant overall difference ($p = 0.0244$). However, none of the pairwise comparisons between clinical entities reached the corrected significance threshold.

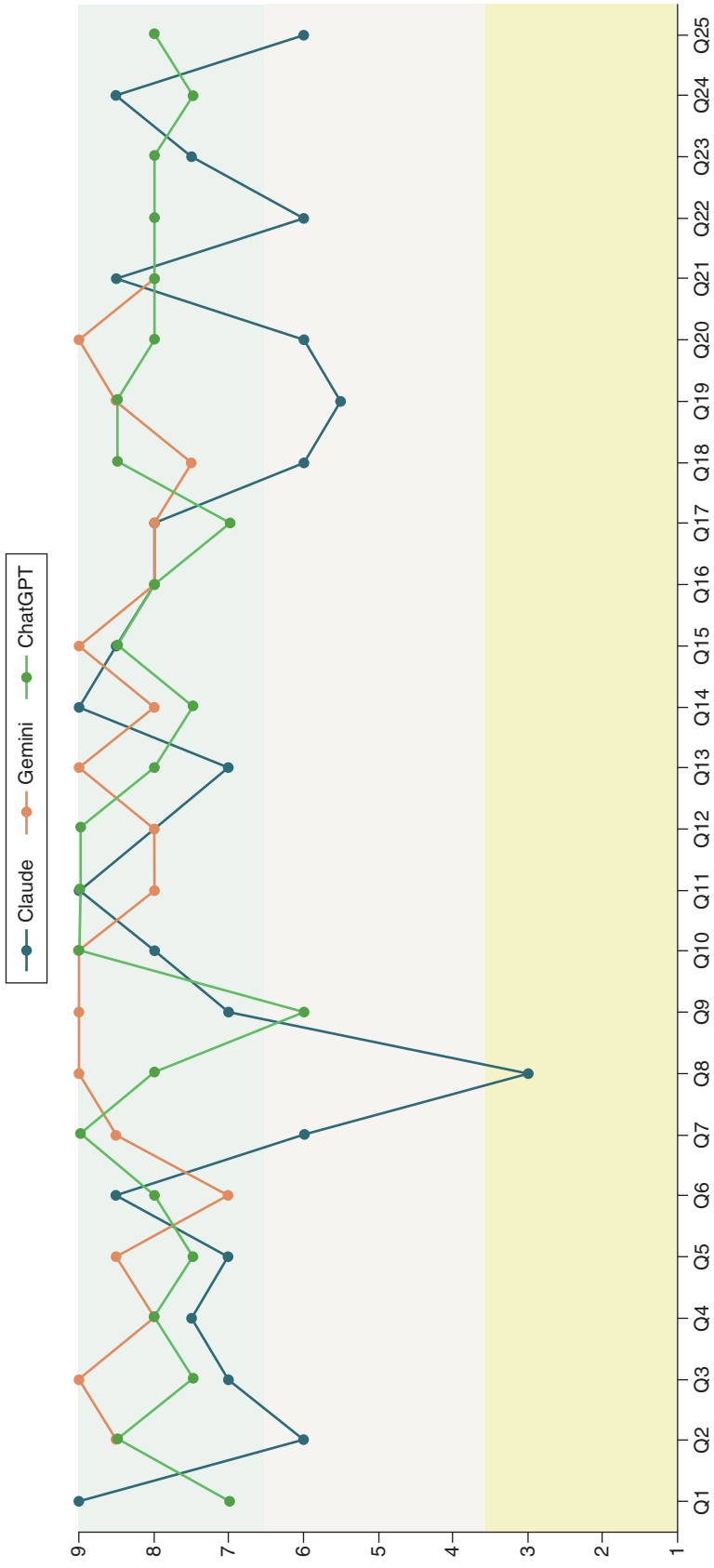


Figure 5. Median scores assigned by panelists for each response and large language model in upper tract urothelial carcinoma. Median scores on a 9-point scale are shown for each clinical question and large language model (Claude, Gemini and ChatGPT). Score ranges were defined as: 1–3 = disagreement, 4–6 = no consensus, 7–9 = agreement. Clinical questions are discrete and not sequentially ordered; connecting lines are used only to facilitate visual comparison among models and do not represent trends across questions. The number of panelists, median score and interquartile range for each question and model are reported in Supplementary Table 6.

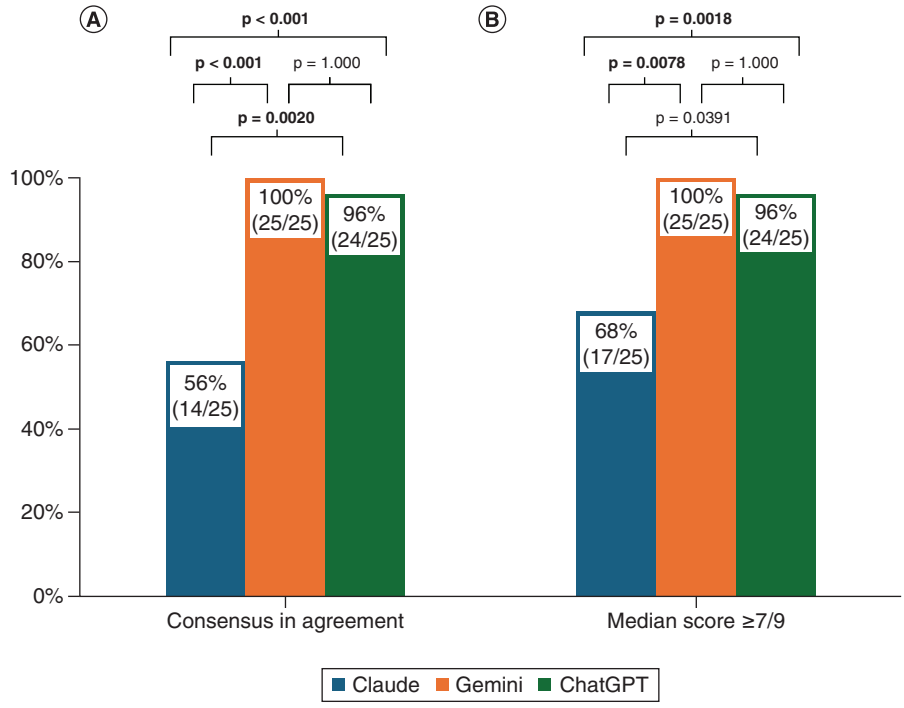


Figure 6. Comparison of expert panel agreement across large language models in upper tract urothelial carcinoma. **(A)** Proportion of consensus in agreement and **(B)** high-scoring items (median ≥7/9) across 25 responses for each large language model in upper tract urothelial carcinoma. Pairwise comparisons were considered statistically significant at $p < 0.0167$ after Bonferroni correction.

Discussion

This study evaluated expert panel agreement with responses generated by Claude, Gemini and ChatGPT to clinical questions based on the supplied EAU guidelines for NMIBC, MIBC and UTUC. Using a modified Delphi methodology, multidisciplinary expert panels independently assessed the degree of agreement with each response. Gemini and ChatGPT responses frequently achieved consensus in agreement, whereas the corresponding proportion was lower for Claude, particularly in UTUC. The consistently high levels of expert panel agreement observed for Gemini and ChatGPT across the three clinical entities support the potential value of these models for guideline interpretation. These findings should nevertheless be interpreted within the study design, which assessed expert agreement rather than independently establishing factual accuracy, clinical safety or real-world clinical utility.

LLMs are being more frequently used as decision-support tools in oncology [8,18]. In UC management, LLMs and AI-driven tools may support a wide range of healthcare professionals beyond clinicians, including radiologists, pathologists, nurses and healthcare administrators. These technologies may assist in critical tasks such as diagnostic imaging interpretation, detailed pathology report analysis, personalized treatment planning, patient monitoring and enhancing communication across multidisciplinary teams [13,19–21]. Although these technologies demonstrate significant potential to enhance evidence-based decision-making, their clinical accuracy and reliability still vary, particularly in complex or changing situations, emphasizing the need for caution before implementing them in patient care [8,22].

In our study evaluating responses generated by LLMs to guideline-based questions on NMIBC, MIBC and UTUC, expert panel agreement varied across models and clinical entities. For NMIBC and MIBC, overall agreement among experts was high, generally above 88% for all LLMs, and there were no significant pairwise differences among models. In contrast, in the context of UTUC, Claude's consensus in agreement rate was substantially lower (56%, 14/25 responses), compared with nearly complete consensus in agreement for Gemini and ChatGPT (100%, 25/25 and 96%, 24/25 responses, respectively). Panelist feedback frequently cited incomplete explanations or a lack of sufficient clinical nuance in specific responses, especially for more complex or atypical clinical scenarios. Moreover, the proportion of responses evaluated with a median score of 7 or higher out of 9 was notably lower for Claude (68%, 17/25 responses) than for Gemini (100%, 25/25 responses) and ChatGPT (96%, 24/25 responses).

These findings indicate variation in expert panel agreement but do not establish that the responses were factually correct, complete, clinically actionable or safe. Qualitative feedback also included comments that some Claude responses were insufficiently focused, too brief or overly elaborate, suggesting that response style and level of detail, in addition to factual concerns, may have influenced panelist ratings. However, because the comments were not collected using predefined categories, we cannot reliably determine whether these patterns reflected overly cautious or hedged responses rather than differences in the substantive content of the responses.

A notable challenge in evaluating LLMs in medical contexts is the lack of standardized methodology across studies [8]. As recent literature highlights, evaluation approaches vary widely, from multiple-choice assessments and open-ended clinical prompts to expert ratings and consensus methods, making direct comparisons difficult [8]. In our current study, we applied a Delphi methodology involving a multidisciplinary expert panel including urologists, medical oncologists, radiation oncologists, nuclear medicine physicians, pathologists, radiologists and hospital pharmacists, enabling structured, iterative consensus building and rich qualitative feedback. The number and background of experts were similar to our previous study, which used a simpler direct rating system with experts independently scoring responses on a 5-point Likert scale [14]. Interestingly, the previous study found that rephrased prompts did not consistently improve expert ratings and that simple prompts received higher ratings in most comparisons. Claude with simple prompts achieved the highest overall proportion of optimal ratings [14]. In contrast, the current study used only simple prompts and observed that Gemini and ChatGPT, based on later model versions released in 2025, had higher proportions of responses achieving consensus in agreement and median scores ≥ 7 than Claude, whose version dates from late 2024. Because the studies evaluated different model versions at different time points, this reversal should not be interpreted as a stable ranking of vendors or model families. It may reflect version-specific updates together with differences in evaluation methodology, and confirmation would require contemporaneous testing under identical prompting conditions with repeated sampling. Additionally, the limited and less detailed clinical evidence for rarer tumors such as UTUC [5] may also contribute to variation in expert panel agreement. Together, these findings underscore the importance of rigorous multidisciplinary evaluation and continuous validation when interpreting expert ratings of responses generated by rapidly evolving LLMs.

Several recent studies provide important context for interpreting our findings in both general and urological oncology. Longwell *et al.* evaluated multiple LLMs with standardized oncology society exam questions, finding the top proprietary LLM answered 85% correctly but with some clinically harmful errors, highlighting the need for expert oversight [11]. Eckrich *et al.* compared LLM output to urology consultants using case-based questions, noting that while LLMs generated coherent answers, they had lower medical adequacy and a higher rate of potentially hazardous misinformation than experts [10]. Similarly, 't Hoen *et al.* assessed LLMs' ability to update pediatric urology guidelines and found that, despite scientific accuracy in many cases, consistency and reliability levels were insufficient to replace expert review [12]. These findings collectively emphasize the promise of LLMs in oncology while underscoring the necessity of expert supervision and continuous evaluation before clinical deployment. In the specific context of BC management, recent work by Li *et al.* comprehensively evaluated six leading LLMs on clinical questions spanning diagnosis, treatment, prognosis and follow-up. Claude-3.5-Sonnet demonstrated the highest baseline accuracy, while iterative fine-tuning enabled GPT-3.5-Turbo to reach perfect accuracy [13]. These results highlight both the potential and current limitations of LLMs in providing reliable, guideline-concordant support, emphasizing the ongoing need for expert validation, sophisticated training methods, and integration of continuous clinical feedback to optimize AI-driven decision support in UC care.

This study has several limitations that should be acknowledged when interpreting its findings. It focused on three urothelial carcinoma subtypes (NMIBC, MIBC and UTUC) using a relatively limited set of 25 predefined clinical questions per clinical entity, which may reduce the generalizability of results to other clinical scenarios or cancers. The evaluation employed a two-round Delphi process with multidisciplinary expert panels including urologists, medical oncologists, radiation oncologists, nuclear medicine physicians, pathologists, radiologists and hospital pharmacists. While this approach facilitates structured consensus building and captures rich qualitative feedback, it inherently involves some subjectivity and potential bias related to panel composition and individual interpretations. Because members of the specialized tumor boards contributed to question development and subsequently evaluated the LLM responses, this overlap of roles may have introduced incorporation bias. However, panelists were blinded to model identity, and the plain language descriptions of figures and tables were reviewed by independent tumor board members. 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 considered when interpreting the results. The limited number of questions and panelists, together with variable voting participation, may also have reduced statistical power, particularly for pairwise model comparisons. Therefore, the absence of statistically significant differences should not be interpreted as evidence of equivalence between models. At the same time, the consistently high expert panel agreement observed for Gemini and ChatGPT provides descriptive evidence of similarly high levels of agreement for these two models under the specific study conditions. Furthermore, rapid and frequent updates of LLM architectures mean that model performance may vary over time. The evaluated models were also released at different times, from October 2024 to May 2025. Consequently, the observed differences should be interpreted as comparisons between the specific versions tested and cannot be attributed solely to vendors or assumed to represent stable differences between model families. Only one response was generated per question and model, so reproducibility across repeated runs could not be assessed, while the use of English guideline materials with Spanish questions may limit generalizability to other language settings. Because the models were provided with the relevant EAU guideline content, including figures and tables converted into plain language textual summaries, the study primarily evaluated their ability to interpret and apply supplied guideline information under controlled conditions. The study did not include a predefined reference standard for independently assessing factual accuracy, completeness, clinical actionability or safety, and the qualitative comments were not collected using standardized categories that would permit reliable retrospective quantification of these outcomes. Therefore, consensus in agreement should not be interpreted as demonstrating that a response was fully correct, complete, clinically actionable or safe, or as evidence of performance in real clinical settings. Nevertheless, the structured preparation of the guideline content, combined with the multidisciplinary Delphi evaluation, strengthens the consistency and clinical relevance of the assessment. Continued evaluation and cautious interpretation of LLM outputs remain essential as these models evolve.

In conclusion, when provided with the 2025 EAU guidelines, including figures and tables converted into plain language textual summaries, the evaluated LLMs frequently generated responses that achieved multidisciplinary expert panel agreement. Agreement was particularly high for Gemini and ChatGPT across the three urothelial carcinoma settings evaluated, supporting the potential value of these models for guideline interpretation. Although the limited number of questions and variable panel participation warrant caution when interpreting model-to-model comparisons, the consistently high expert panel agreement observed for Gemini and ChatGPT provides a strong rationale for further evaluation using broader question sets and in real-world clinical settings. Such studies are needed to establish the clinical utility and safety of these tools as clinical decision-support systems.

Summary points

- This study evaluated Claude 3.5 Haiku, Gemini 2.5 Flash and GPT-4.1-mini using 25 guideline-based questions for each of three clinical entities: non-muscle-invasive bladder cancer, muscle-invasive bladder cancer and upper tract urothelial carcinoma.
- The models received the same 2025 European Association of Urology guideline materials, with figures and tables converted into plain language text.
- Model responses were assessed by multidisciplinary expert panels through a two-round modified Delphi process.
- Across all 75 questions, consensus in agreement was achieved for 78.7% of Claude responses, 100% of Gemini responses, and 96% of ChatGPT responses.
- No significant pairwise differences in consensus in agreement were identified among the models for non-muscle-invasive or muscle-invasive bladder cancer.
- For upper tract urothelial carcinoma, significantly fewer Claude responses achieved consensus in agreement than Gemini or ChatGPT responses.
- Across all questions, median expert panel scores $\geq$7 were achieved for 84% of Claude responses, 100% of Gemini responses and 98.7% of ChatGPT responses.
- The findings apply to the specific model versions and controlled guideline-based conditions evaluated and should be interpreted cautiously given the limited question set and variable panel participation.
- The consistently high expert panel agreement observed for Gemini and ChatGPT supports further evaluation of guideline-conditioned LLMs to establish their clinical utility and safety.

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-2026-0079

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 Dr 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 competing interests or relevant affiliations with any organization or entity with the subject matter or materials discussed in the manuscript. This includes employment, consultancies, honoraria, stock ownership or options, expert testimony, grants or 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 75 questions is 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. Dyrskjøt L, Hansel DE, Efstathiou JA *et al.* Bladder cancer. *Nat. Rev. Dis. Primers* 9(1), 58 (2023).
2. Escobar D, Wang C, Suboc N, D'Souza A, Tulpule V. Diagnosis and management of upper tract urothelial carcinoma: a review. *Cancers (Basel)* 17(15), 2467 (2025).
3. European Association of Urology. EAU guidelines on non-muscle-invasive bladder cancer. (Accessed 19 May 2025). https://uroweb.org/guidelines/non-muscle-invasive-bladder-cancer
4. European Association of Urology. EAU guidelines on muscle-invasive and metastatic bladder cancer. (Accessed 19 May 2025). https://uroweb.org/guidelines/muscle-invasive-and-metastatic-bladder-cancer
5. European Association of Urology. EAU guidelines on upper urinary tract urothelial cell carcinoma. (Accessed 19 May 2025). https://uroweb.org/guidelines/upper-urinary-tract-urothelial-cell-carcinoma
6. Ferlay J, Ervik M, Lam F *et al.* Bladder 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/30-bladder-fact-sheet.pdf
7. Fernandez Aparicio T, Galan Llopis JA, Cansino Alcaide R *et al.* Incidence of upper tract urothelial carcinoma in Spain. *Actas Urol. Esp.* 44(7), 512–518 (2020).
8. Carl N, Schramm F, Haggenmüller S *et al.* Large language model use in clinical oncology. *NPJ Precis. Oncol.* 8(1), 240 (2024).

- **Systematic review and meta-analysis of 34 studies highlighting substantial heterogeneity in large language model (LLM) performance and evaluation methodologies in clinical oncology, and the need for standardized reporting.**

9. 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).

10. Eckrich J, Ellinger J, Cox A *et al.* Urology consultants versus large language models: potentials and hazards for medical advice in urology. *BJUI Compass* 5(5), 438–444 (2024).

- **Blinded comparison of responses to 45 urological case-based questions from three LLMs and urology consultants, showing lower medical adequacy and more potential misinformation hazards for LLM-generated responses despite their generally good coherence and comprehensibility.**

11. Longwell JB, Hirsch I, Binder F *et al.* Performance of large language models on medical oncology examination questions. *JAMA Netw. Open* 7(6), e2417641 (2024).

12. 't Hoen LA, van Uitert A, Bussmann M *et al.* Annual updates of the European Association of Urology–European Society for Pediatric Urology (EAU–ESPU) paediatric urology guidelines: are large-language models (LLM) better than the usual structured methodology? *J. Pediatr. Urol.* 21(6), 1643–1649 (2025).

•• **Evaluation of 60 repeated conversations testing three LLMs for updating two EAU–ESPU pediatric urology guideline chapters, showing inadequate accuracy, consistency, reliability and ability to incorporate recent literature.**

13. 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.* 39(5), 494–499 (2025).

•• **Comparative evaluation of six LLMs across 100 guideline-based bladder cancer questions in three independent trials, showing substantial differences in baseline accuracy and marked improvement following targeted model optimization.**

14. 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).

•• **Direct methodological precursor evaluating six configurations of Claude, Gemini and ChatGPT across 175 guideline-based questions covering seven major urological cancers, using 10,500 expert ratings and comparing simple and rephrased prompts.**

15. Anthropic. Models overview. Claude Platform Docs. (Accessed 23 June 2025). https://platform.claude.com/docs/en/about-claude/models/overview

16. Google AI for Developers. Gemini 2.5 Flash. Gemini API. (Accessed 23 June 2025). https://ai.google.dev/gemini-api/docs/models/gemini-2.5-flash

17. OpenAI. GPT-4.1 mini model. OpenAI API. (Accessed 23 June 2025). https://developers.openai.com/api/docs/models/gpt-4.1-mini

18. Zitu MM, Le TD, Duong T *et al.* Large language models in cancer: potentials, risks, and safeguards. *BJR Artif. Intell.* 2(1), ubae019 (2025).

19. Maggi M, Chierigo F, Fallara G *et al.* Shaping the future of personalized therapy in bladder cancer using artificial intelligence. *Eur. Urol. Focus* 12(1), 41–48 (2026).

20. Sun D, Hadjiiski L, Gormley J *et al.* Outcome prediction using multimodal information: integrating large language model-extracted clinical information and image analysis. *Cancers (Basel)* 16(13), 2402 (2024).

21. Duwe G, Mercier D, Kauth V *et al.* Development of an artificial intelligence-generated, explainable treatment recommendation system for urothelial carcinoma and renal cell carcinoma to support multidisciplinary cancer conferences. *Eur. J. Cancer* 220, 115367 (2025).

22. 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(9), 2613–2622 (2024).

•• **Large-scale evaluation using 2400 real patient cases in a simulated clinical workflow, showing that current LLMs underperformed physicians, did not reliably follow diagnostic or treatment guidelines, and were not ready for autonomous clinical decision-making.**