The New Journal of Urology
eISSN 3023-6940
doi: https://doi.org/10.33719/nju1968495

RESEARCH ARTICLE

Often Correct, Not Reliably Correct: A Blinded Audit of GPT-5.2 as Guideline-Decision Support for a Rare Cancer (Penile Squamous Cell Carcinoma)

Ahmet Şanlı [1 *], Serhat Göçer [2], Hasan Samet Güngör [3]

[1] Department of Urology, Karamanoğlu Mehmetbey University Training and Research Hospital, Karaman, Türkiye
[2] Department of Urology, Tekirdağ City Hospital, Tekirdağ, Türkiye
[3] Department of Urology, Health Sciences University, Ümraniye Training and Research Hospital, Istanbul, Türkiye

Submitted: *2026-06-11*
Accepted: *2026-08-24*

Corresponding Author;
Ahmet Şanlı, MD
Address: Department of Urology, Karamanoğlu Mehmetbey University Training and Research Hospital, Şehit Ömer Halis Demir Street 7/1, 70100 Karaman, Türkiye
E-mail: drasanli07@gmail.com

ORCID
A.Ş. 0000-0001-8237-1072
S.G. 0000-0001-6982-4663
H.S.G. 0000-0002-7632-1199

Abstract

Objective: Large language models (LLMs) are increasingly used as guideline-decision support in oncology, but most validation has occurred in common cancers, where clinicians can usually recognize errors. The safety-critical setting is the opposite — rare cancers, where local expertise is limited, and errors may go unrecognized. Whether common-cancer concordance reflects the reliability needed there is unclear.

Methods: We benchmarked GPT-5.2 against the EAU-ASCO 2025 Penile Cancer Guidelines using 100 guideline-derived questions across five domains and three zero-shot prompting strategies, yielding 300 responses. Two experienced urologists, blinded to prompt identity, scored concordance on a 3-point scale and classified errors as hallucination, omission, temporal, or reasoning. The primary analysis was a mixed-effects ordinal logistic regression with question as a random intercept; a cluster-robust logistic regression served as sensitivity. Reporting followed TRIPOD-LLM.

Results: GPT-5.2 was fully concordant in 236 of 300 responses (78.7%; 95% confidence interval, 73.7–82.9; weighted κ=0.914). Prompt strategy did not improve performance (role-play +3 percentage points; 95% credible interval, −2 to +9). The domain effect was not robust once between-question clustering was modeled. Among 64 non-concordant responses, omission predominated (47.7%) and reasoning errors followed (37.5%); in treatment, omission accounted for 68.0% of errors, including loss of guideline-specified detail. Four questions were uniformly discordant across all three prompts.

Conclusion: Headline concordance overstated clinical reliability. The clinically informative finding was not the overall accuracy but the structure of failure — omission of decision-critical treatment detail and stable errors. Prompt engineering did not rescue performance. Current LLM output may assist orientation in rare-cancer decision pathways but is not ready as stand-alone guideline-decision support.

Keywords: artificial intelligence, natural language processing, penile neoplasms, practice guidelines as topic, clinical decision-making, reproducibility of results

Cite; Sanli A, Gocer S, Gungor HS. Often Correct, Not Reliably Correct: A Blinded Audit of GPT-5.2 as Guideline-Decision Support for a Rare Cancer (Penile Squamous Cell Carcinoma). New J Urol. 2026;21(3):151–161. **doi:** https://doi.org/10.33719/nju1968495

This work is licensed under a Creative Commons Attribution 4.0 (CC-BY) International License. © *Copyright remains with the authors.*

INTRODUCTION

Large language models (LLMs) are increasingly used as guideline-decision support, often outside the conditions in which they were benchmarked (1). In oncology, that benchmarking has concentrated on common cancers, where the clinicians most likely to query a model are also the most likely to recognize an error (2). The deployment setting that matters most for safety is the opposite. An uncommon diagnosis, managed by clinicians with limited disease-specific experience, is precisely where decision support is most clinically valuable and where its errors are most likely to go unrecognized. Whether the concordance rates reported in common-cancer benchmarks generalize to that setting — and whether they reflect the form of reliability needed there — is unclear.

Penile squamous cell carcinoma presents a different problem. It is rare, with a global incidence of 0.80 per 100,000 males and fewer than 3,100 new cases annually across the European Union (3–5). Many urologists encounter it infrequently, particularly outside referral centers. In this setting, adherence to guideline-based management is especially important and has been associated with better outcomes (6). The same setting also creates the conditions in which a clinician may be tempted to seek rapid support from an LLM (7–8).

Previous LLM benchmarking studies in oncology, including evaluations of urological cancers, have reported concordance estimates generally ranging from 77% to 89% (9–11). A recent meta-analysis across 15 cancer types reported a pooled accuracy of 76.2% (12). Penile cancer was absent from those evaluations. As a result, it remains unclear whether performance observed in common cancers generalizes to a rare malignancy in which LLM performance may be less predictable and clinically important errors may carry greater consequences (13).

We therefore evaluated GPT-5.2 against the European Association of Urology-American Society of Clinical Oncology (EAU-ASCO) 2025 Penile Cancer Guidelines (14) using 100 guideline-derived clinical questions, three prompting strategies, and blinded expert review. In addition to concordance, we examined error structure, cross-prompt consistency, and question-level heterogeneity to assess not only whether the model was right, but also where it failed and how stable those failures were. The study was designed and reported in accordance with the Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis-Large Language Model (TRIPOD-LLM) statement (15) (Supplementary Table S1).

MATERIALS AND METHODS

Study Design and Benchmark Construction

We conducted a cross-sectional benchmarking study of GPT-5.2 (OpenAI, San Francisco, CA, USA) against the EAU-ASCO 2025 Penile Cancer Guidelines (14). The benchmark consisted of 100 clinical questions derived from the full guideline text. Questions were developed by a urologist with experience in urological oncology (A.Ş.) and were restricted to statements with objectively verifiable answers. Gold-standard responses were extracted directly from the guideline and verified by section and page number. The EAU-ASCO 2025 Penile Cancer Guidelines used as the reference standard is a reprint of the 2023 edition; potential overlap with the model's training data was not corrected for and is addressed in the Limitations.

To ensure broad clinical coverage, the number of questions per domain was allocated according to the size of the corresponding guideline section: epidemiology (n = 15), tumor, node, metastasis (TNM) classification (n = 20), diagnosis (n = 20), treatment (n = 30), and follow-up (n = 15). Questions were balanced by type, with 50 definition/classification items and 50 action-oriented items. They were also categorized by recommendation strength: 37 strong, 14 weak, and 49 not graded. Two independent urologists reviewed all questions and gold-standard responses for fidelity to the source guideline and for clinical relevance. Discrepancies were resolved by consensus. Question clarity was pilot tested before formal evaluation. The complete question set with gold-standard answers is provided in Supplementary Table S2.

Prompting Strategy and Testing Conditions

Each question was submitted to GPT-5.2 under three zero-shot prompting strategies: baseline, role-play, and chain-of-thought (Supplementary Figure S1). The baseline condition presented the question without modification. The role-play condition instructed the model to respond as a senior urologic oncologist and EAU guideline panel member. The

chain-of-thought condition instructed the model to analyze the scenario step by step (1,16). Full prompt texts are provided in Supplementary Table S3.

All 300 queries were submitted through the ChatGPT web interface between February 9, 2026, and February 19, 2026. Testing was performed under controlled conditions. A new account was used, memory was disabled, web search was turned off, no system-level or developer-level instructions were applied, and each question was submitted in a separate chat (Supplementary Figure S1). Because the web interface does not permit direct control of generation settings, default model parameters were used throughout. All 300 raw LLM responses with corresponding reviewer scores are available in Supplementary Table S4.

Blinded Evaluation Framework

Responses were assessed using a 3-tier framework (Supplementary Table S5). Tier 1 captured guideline concordance on a 3-point scale: concordant (2), partially concordant (1), or discordant (0). Tier 2 classified non-concordant responses into predefined error categories: hallucination, omission, temporal error, and reasoning error (17,18). Tier 3 assessed cross-prompt consistency.

Two experienced urologists independently scored all 300 responses. Responses were randomized separately for each reviewer, and prompt identity was masked. A 15-question calibration set was scored jointly by both reviewers to harmonize scoring criteria and resolve interpretive differences. The remaining 85 questions were scored independently. Consensus scores were obtained by rounding the mean of the two reviewers. Both reviewers independently classified error types for all responses with a consensus concordance score of 0 or 1.

Statistical Analysis

Inter-rater agreement for concordance scores was summarized with quadratic weighted Cohen's kappa (19). A 95% confidence interval was obtained by bootstrapping over questions. Each question produced three correlated responses. Concordance score (0,1,2) was therefore modeled with a mixed-effects ordinal (proportional-odds) logistic regression, with domain and prompt strategy as fixed effects and question as a random intercept. Most questions were scored identically across the three prompts. This near-uniformity produced extreme between-question separation, and maximum-likelihood estimation of the random-intercept variance became unstable at the boundary. We therefore fitted a Bayesian hierarchical ordinal model with weakly-informative priors and designated it the primary analysis. That designation was made after the data were seen. Priors were Normal (0, 2.5) for the fixed effects and half-Normal (2.5) for the random-intercept standard deviation. The model was fitted by Hamiltonian Monte Carlo with four chains and 2000 post-warmup draws each. Convergence was confirmed by R-hat and effective sample size.

Fixed effects were summarized as posterior odds ratios with 95% credible intervals. The overall contributions of domain and prompt were assessed by leave-one-out cross-validation and by likelihood-ratio tests. Both remain better calibrated than Wald tests under separation. The intraclass correlation was computed on the latent scale as $\sigma^2/(\sigma^2+\pi^2/3)$. The random-intercept variance is weakly identified under separation. We therefore examined the sensitivity of every reported quantity to the prior across five specifications (Supplementary Table S6). The proportional-odds assumption was examined by comparing coefficients across the two cumulative cut-points. Maximum-likelihood estimation of the same model, with Wald tests, was retained as a sensitivity analysis. For the binary contrast of full concordance versus not, domain odds ratios were estimated with Firth penalization, because the reference domain sat near the ceiling.

For the prompt comparison, we report the difference in full-concordance probability in percentage points, with its credible interval. We also state the improvement the data can and cannot exclude. The study was not sized to establish equivalence, and none is claimed. Supportive analyses included descriptive concordance rates with Wilson intervals, the Friedman test across prompts, the Kruskal-Wallis test across domains with Bonferroni-corrected pairwise Mann-Whitney U tests, and Spearman correlation between concordance and consistency. Because most questions were scored identically across prompts, the Friedman ranks were heavily tied, and the comparison rests on the questions that varied. Agreement on error type was summarized with Cohen's kappa, a prevalence-adjusted kappa, and category-specific agreement with bootstrap intervals. All tests were 2-sided. Analyses used

Python (SciPy, scikit-learn, statsmodels) and PyMC with ArviZ. The analysis code is available from the corresponding author on reasonable request.

No ethical approval was required because the study did not involve human participants, patient-level identifiers, or animal subjects.

RESULTS

Agreement between the two blinded reviewers was excellent. The quadratic weighted kappa for concordance scoring was 0.914 (95% CI, 0.835 to 0.967), with 94.0% exact agreement across 300 responses and no 2-point disagreements (Table 1).

Of the 300 responses, 236 (78.7%; 95% confidence interval [CI], 73.7% to 82.9%) were fully concordant with the guideline, 41 (13.7%) were partially concordant, and 23 (7.7%) were discordant (Table 1; Supplementary Table S7). Thus, roughly four of five outputs aligned fully with guideline recommendations, whereas about one in thirteen contradicted them.

The prompt strategy did not materially change performance. Full concordance was 77% with baseline prompting, 80% with role-play, and 79% with chain-of-thought. In the primary Bayesian model, the role-play and chain-of-thought odds ratios both had 95% credible intervals that included 1, and cross-validation did not favor a model with prompt. The result was unchanged in supportive analyses, including the Friedman test (χ^2=1.23, p = 0.54) and the binary cluster-robust sensitivity model (Wald χ^2=0.99, p = 0.61) (Figure 1B).

Concordance varied across domains at the descriptive level. Follow-up had the highest full concordance rate at 93.3%, whereas epidemiology had the lowest at 71.1% and also the highest discordance rate at 20.0%. This variation did not survive the structure of the data. Eighty-one of the 100 questions were scored identically across all three prompts. Most of the variance therefore lay between questions, not between domains or prompts. The intraclass correlation was high (0.87; 95% credible interval, 0.78 to 0.93). The random-intercept standard deviation was large and weakly identified (posterior median ≈4.7 on the logit scale; prior-sensitive, Supplementary Table S6). In the primary Bayesian model, every domain odds ratio had a 95% credible interval that included 1. The overall domain effect was not significant by likelihood-ratio test (χ^2=9.19, df=4, p = 0.06), and cross-validation did not distinguish the model with domain from the model without it. Domain is a between-question characteristic, and only 19 questions varied across prompts. The study had limited power to resolve a domain effect (Table 2, Panel A; Figure 1A; Supplementary Table S8; Supplementary Figure S2).

The response-level analyses pointed the other way. The Kruskal-Wallis test was nominally significant (H=13.95, df=4, p = 0.007), and a Firth-penalized logistic model placed the odds of full concordance below follow-up for epidemiology, diagnosis, and treatment (Table 2, Panel B). That model treats the 300 responses as independent. It ignores the three responses clustered within each question, and it overstates the precision of a between-question contrast. Once the clustering was restored, the domain effect did not survive. The domain gradient is suggestive, not established.

Table 1. Overall guideline concordance and inter-rater reliability by prompt strategy

	n	Concordant (score 2)	Partial (score 1)	Discordant (score 0)	Mean score	κw (95% CI)	% Agreement
Prompt A (Baseline)	100	77 (77.0%)	14 (14.0%)	9 (9.0%)	1.68	0.909 (0.823–0.967)	93.0%
Prompt B (Role-play)	100	80 (80.0%)	13 (13.0%)	7 (7.0%)	1.73	0.907 (0.812–0.972)	94.0%
Prompt C (Chain-of-thought)	100	79 (79.0%)	14 (14.0%)	7 (7.0%)	1.72	0.926 (0.836–0.984)	95.0%
Overall	300	236 (78.7%)	41 (13.7%)	23 (7.7%)	1.71	0.914 (0.835–0.967)	94.0%

Concordance scores represent consensus between two blinded reviewers. κw = quadratic weighted Cohen's kappa. Friedman test: χ^2=1.23, p = 0.54 (no significant difference between prompt strategies).

Weighted-kappa 95% confidence intervals were obtained by bootstrapping over questions. Friedman test df=2; because most questions were scored identically across prompts, the ranks were heavily tied.

Question type was not associated with concordance. Action-oriented questions did not perform differently from definition/classification items in either the descriptive comparison (Mann-Whitney p = 0.41) or the cluster-robust model (OR=1.24; 95% CI, 0.52–2.97; p = 0.62). The recommendation strength also did not explain model performance. Strong recommendations had a lower mean concordance than weak recommendations (1.61 versus 1.82), but the difference was not statistically significant in either the descriptive analysis (p = 0.20) or the cluster-robust model (OR for weak versus strong, 2.19; 95% CI, 0.50–9.52; p = 0.30; Supplementary Table S9).

Among the 64 consensus non-concordant responses, omission and reasoning failures predominated. Both reviewers identified similar error distributions. Reviewer 1 classified 46.9% as omission, 37.5% as reasoning errors, and 15.6% as hallucinations. Reviewer 2 classified 48.4% as omission, 37.5% as reasoning errors, and 14.1% as hallucinations. Averaging across reviewers, omission accounted for 47.7% of all errors, reasoning for 37.5%, and hallucination for 14.8% (Table 3; Figure 2A).

Agreement on error type was substantial overall (Cohen's kappa=0.795; prevalence-adjusted kappa=0.81), but it varied by category. It was least reliable for hallucination (category-specific kappa=0.57; 95% CI, 0.24 to 0.84). Agreement was lowest in follow-up. There, the three non-concordant responses came from one question, which one reviewer scored as omission and the other as hallucination. That single cell should not be read as a stable estimate.

Table 2. Primary and sensitivity regression analyses for guideline concordance

Panel A. Mixed-effects ordinal logistic regression*					
Fixed effect	χ^2 (LRT)	**df**	***p*-value**		
Domain (overall)	9.19	4	0.06		
Prompt strategy (overall)	1.60	2	0.45		
Random-intercept SD (σ), posterior median (95% CrI)	4.7 (3.4–6.6)				
Intraclass correlation (95% CrI)	0.87 (0.78–0.93)				
Panel B. Binary cluster-robust logistic regression (sensitivity analysis)†					
Variable	**Full concordance (%)**	**OR**	**95% CI**	***p*-value**	**Wald test‡**
Domain					χ^2=4.84, p = 0.30
Follow-up (ref.)	93.3	1.00	—	—	
Epidemiology	71.1	0.20	0.05–0.64	—	
TNM	88.3	0.59	0.14–2.11	—	
Diagnosis	73.3	0.22	0.06–0.69	—	
Treatment	72.2	0.21	0.05–0.62	—	
Prompt strategy					χ^2=0.99, p = 0.61
Baseline (ref.)	77.0	1.00	—	—	
Role-play	80.0	1.19	0.84–1.70	0.32	
Chain-of-thought	79.0	1.12	0.75–1.67	0.57	

*Mixed-effects ordinal logistic regression (proportional odds) with concordance score (0,1,2) as outcome, domain and prompt as fixed effects, question ID as random intercept. In the primary analysis the model was fitted by Bayesian estimation with weakly-informative priors; a maximum-likelihood fit with adaptive Gauss-Hermite quadrature (25 nodes) served as a sensitivity analysis. Individual domain odds ratios are not reported because the large, weakly-identified random-intercept variance renders conditional odds ratios unstable. Overall effects are assessed by likelihood-ratio χ^2 test and by leave-one-out cross-validation; Panel A gives the posterior random-intercept standard

deviation (prior-sensitive; Supplementary Table S6) and the intraclass correlation. †Binary logistic regression (fully concordant [score 2] vs not fully concordant [score 0 or 1]) with cluster-robust (sandwich) standard errors, clustered by question ID. Domain odds ratios are Firth-penalized to remain finite under the near-ceiling reference domain; this marginal model treats the 300 responses as independent, whereas every domain 95% credible interval in the primary clustered model included 1. OR <1 indicates lower odds of full concordance relative to the reference category. ‡Overall Wald chi-square test. Descriptive concordance rates by domain and prompt are provided in Supplementary Table S8.

Table 3. Error taxonomy by clinical domain — dual-reviewer classification (n = 64 consensus non-concordant responses)

			Reviewer 1			Reviewer 2		
Clinical domain	**n**		**Hallucination n (%)**	**Omission n (%)**	**Reasoning n (%)**	**Hallucination n (%)**	**Omission n (%)**	**Reasoning n (%)**
Epidemiology	13		3 (23.1%)	4 (30.8%)	6 (46.2%)	2 (15.4%)	4 (30.8%)	7 (53.8%)
TNM	7		1 (14.3%)	3 (42.9%)	3 (42.9%)	1 (14.3%)	3 (42.9%)	3 (42.9%)
Diagnosis	16		3 (18.8%)	5 (31.2%)	8 (50.0%)	3 (18.8%)	5 (31.2%)	8 (50.0%)
Treatment	25		3 (12.0%)	15 (60.0%)	7 (28.0%)	0 (0.0%)	19 (76.0%)	6 (24.0%)
Follow-up	3		0 (0.0%)	3 (100.0%)	0 (0.0%)	3 (100.0%)	0 (0.0%)	0 (0.0%)
Total	**64**		**10 (15.6%)**	**30 (46.9%)**	**24 (37.5%)**	**9 (14.1%)**	**31 (48.4%)**	**24 (37.5%)**

Error types based on consensus non-concordant responses (concordance score 0 or 1). Both reviewers independently classified errors. Inter-rater agreement on error type: Cohen's kappa=0.795 (87.5% agreement). Hallucination = fabricated clinical data absent from guidelines; Omission = failure to include critical guideline-specified information; Reasoning = correct facts leading to incorrect clinical conclusion. Prevalence-adjusted kappa = 0.81. Category-specific agreement (kappa): hallucination 0.57 (95% CI, 0.24–0.84), omission 0.78, reasoning 0.93. The follow-up cell reflects a single question (ID 96), scored as omission by one reviewer and hallucination by the other, and should not be over-interpreted.

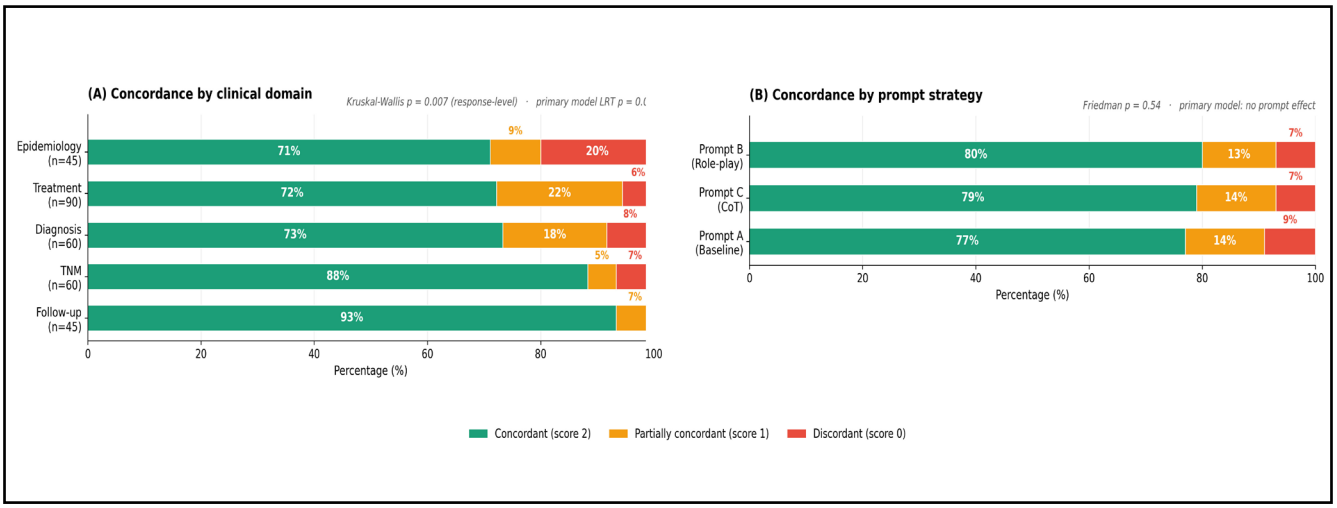


Figure 1. Distribution of guideline concordance scores. Stacked horizontal bar charts showing the proportion of concordant (score 2, green), partially concordant (score 1, amber), and discordant (score 0, red) responses. (A) By clinical domain. Follow-up achieved the highest concordance rate (93.3%, n = 45), whereas Epidemiology had the lowest (71.1%, n = 45) with the highest discordance rate (20.0%). The response-level Kruskal-Wallis test was nominally significant (H=13.95, p = 0.007); the primary Bayesian model showed no robust domain effect (likelihood-ratio χ^2=9.19, df=4, p = 0.06), and the binary cluster-robust sensitivity analysis was consistent (Wald χ^2=4.84, p = 0.30). (B) By prompt strategy. No significant difference was observed among Baseline (77%), Role-play (80%), and Chain-of-thought (79%) strategies (primary Bayesian model: no prompt effect, all 95% credible intervals including 1; Friedman χ^2=1.23, df=2, p = 0.54).

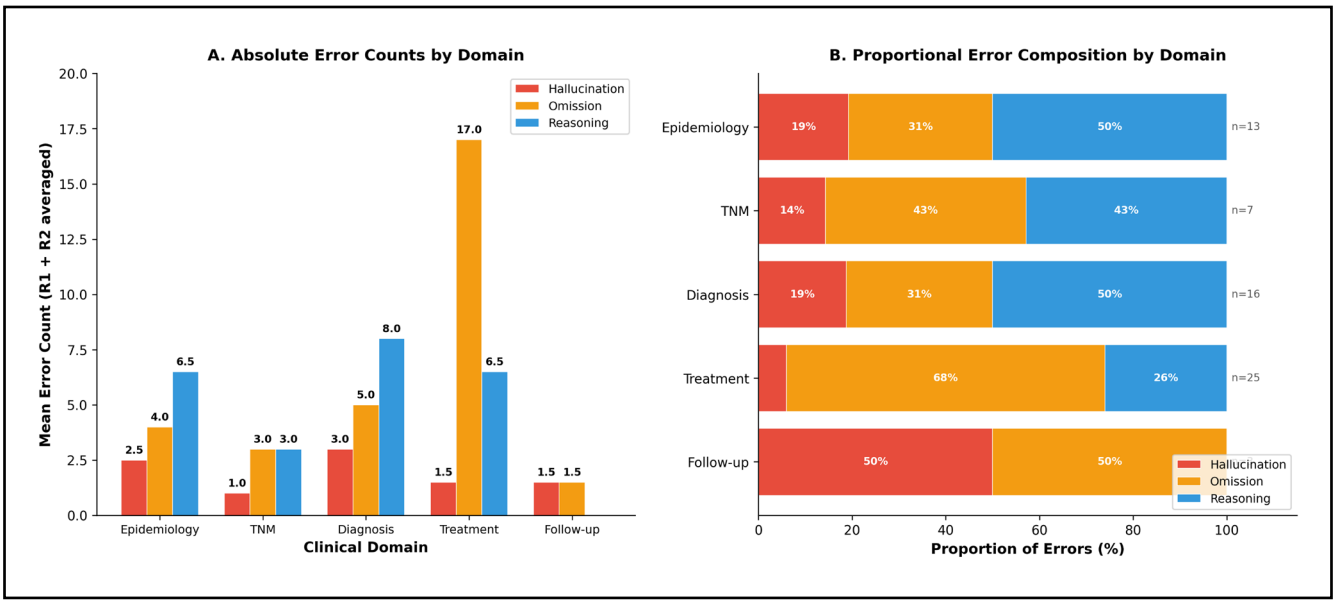


Figure 2. Error type distribution across clinical domains. Based on dual-reviewer assessment (n = 64 consensus non-concordant responses; inter-rater agreement on error type: Cohen's kappa=0.795). Values represent the mean of both reviewers' classifications. (A) Absolute error counts by domain. Treatment generated the highest absolute number of errors (n = 25), predominantly omissions. Epidemiology and diagnosis each showed a predominance of reasoning errors. (B) Proportional error composition within each domain. Reasoning errors were the dominant error type in Epidemiology (50.0%) and Diagnosis (50.0%). Omission predominated in Treatment (68.0%). Hallucination, though less frequent overall (14.8%), was present across domains; in Epidemiology, the model fabricated specific numerical risk estimates not present in the guidelines (19.2% of domain errors). Error categories: hallucination = fabricated clinical data absent from guidelines; omission = failure to include critical guideline-specified information; reasoning = correct facts leading to an incorrect clinical conclusion.

The distribution of errors differed by domain (Figure 2B; Supplementary Table S10). In treatment, omission dominated and accounted for 68.0% of errors. These responses often identified the correct therapeutic direction but omitted critical details specified in the guideline, such as saphenous vein sparing or thresholds for pelvic lymph node dissection. In epidemiology, reasoning errors were most common, accounting for 50.0% of errors, followed by omission (30.8%) and hallucination (19.2%). The hallucinations in this domain involved fabricated numerical estimates for phimosis and smoking risk. Diagnosis also showed a predominance of reasoning errors (50.0%), and TNM classification showed a similar pattern (42.9%).

The most stable feature of the dataset was not the prompt but the question. Eighty-one of 100 questions received identical consensus scores across all three prompting conditions. Seventy-two questions were uniformly concordant, five were uniformly partial, and four were uniformly discordant. The remaining 19 showed mixed performance across prompts.

The uniformly discordant questions were clinically revealing. They included fabricated biopsy specifications, incorrect claims regarding p16 and human papillomavirus (HPV) testing recommendations, and refusal to provide published diagnostic accuracy data. These failures persisted regardless of prompt strategy (Supplementary Tables S11 and S12). Cross-prompt consistency was high overall: 79% of questions received the maximum consistency score. Concordance and consistency were positively correlated (Spearman rho=0.726, $p < 0.001$). Notably, the uniformly discordant questions also received perfect consistency scores, showing that some incorrect answers were not only wrong, but stably wrong.

DISCUSSION

In this rare-cancer benchmark, GPT-5.2 showed moderate agreement with the EAU-ASCO penile cancer guideline, but prompt engineering did not improve performance. The model was often correct. It was not reliably correct. On a subset of questions, it failed reproducibly across all three prompting strategies, and some of the most stable outputs

were incorrect. That distinction matters in precisely the kind of clinical setting in which an LLM is most likely to be used: an uncommon diagnosis, limited operator familiarity, and dependence on guideline-level detail.

The overall concordance of 78.7% sits close to the pooled performance reported across oncology LLM studies and at the lower end of prior reports in urologic cancers (9–12). That result should not be read as reassuring in isolation. First, the benchmark addressed penile cancer, a domain absent from earlier evaluations. Second, the guideline tested here is a reprint of the 2023 version and may already be represented in model training data (14). The observed concordance may therefore approximate best-case performance rather than real-world generalization.

The most consistent negative finding was that prompt engineering did not help. Neither role-play nor chain-of-thought produced a credible improvement. Full concordance was 3 percentage points higher with role-play than with baseline (95% credible interval, −2 to +9), and similar with chain-of-thought. Both intervals include no change, but neither excludes an improvement of roughly eight points. Prompt variation can change wording without reliably improving clinical correctness (20–22). The data are not precise enough to prove prompting inconsequential, and we do not claim equivalence.

The clinically important result was not only whether the model failed, but how it failed. Errors in treatment were dominated by omission. The model often identified the broad treatment direction while omitting the guideline detail that determines whether care is complete. Errors in epidemiology and diagnosis were more often reasoning failures, where apparently correct facts were assembled into incorrect conclusions. Those failures may be harder for clinicians to detect than hallucinations because the output sounds plausible (23). Hallucinations were less common overall, but they were most concerning in epidemiology, where the model generated specific numerical risk estimates not present in the guideline (17,18).

The question-level analyses sharpen this picture. Most questions behaved consistently across prompts, and 81% of them received the same consensus score under all three prompting strategies. Performance, therefore, varied less by prompt phrasing than by the intrinsic content of the question. The model did not merely make isolated errors. On a subset of questions, it failed reproducibly. The high cross-prompt consistency reported in the literature should not be interpreted, on its own, as a marker of safety.

These findings also inform the memorization-versus-reasoning question. If model performance were driven mainly by the retrieval of explicit guideline text, one would expect stronger performance on strong recommendations than on weak ones. That was not observed. Strong recommendations had numerically lower concordance, and the cluster-robust comparison remained non-significant. This does not exclude partial memorization. It does suggest that simple verbatim recovery is not an adequate explanation for model behavior. The presence of fabricated numerical values in epidemiology points in the same direction, consistent with prior reports that LLM responses on epidemiological topics can contain plausible but inaccurate content (24).

This study has limitations. We evaluated one model only. The benchmark contained 100 questions and cannot exhaustively represent a 63-page guideline (14). Different question sets could yield different absolute concordance estimates, although the consistency of prompt, domain, and error patterns suggests that the qualitative findings may be more stable than the exact percentages. The domain effect did not reach statistical significance in clustered analyses, and the wide confidence intervals reflect limited precision with 15 to 30 questions per domain. We also tested zero-shot prompting without retrieval augmentation, which does not represent all deployment scenarios (25). A multi-model extension of this benchmark — incorporating closed-source models from other providers and retrieval-augmented configurations — is the natural next step and would test whether the failure structures we observed are model-specific or representative of current LLM-decision-support behavior in rare-cancer settings.

Two features limit the modeling. Most questions were answered identically across prompts, which produced extreme between-question separation. The random-intercept variance is therefore weakly identified and depends partly on the prior, although the fixed-effect conclusions held across

a range of priors. The primary analysis was also designated after the data were seen. We report it as a post-hoc primary analysis, not a pre-registered one.

CONCLUSION

Even with those limitations, the practical message is clear. Penile cancer is exactly the sort of setting in which clinicians may seek rapid algorithmic support. Our data suggest that current LLM output in this setting is not ready to function as a stand-alone clinical guide. The model was moderately concordant overall, but prompt engineering did not rescue performance; some failures were clinically consequential, and several incorrect outputs were stable across prompts. To conclude, for rare cancers, LLM responses may assist with orientation, but they should not replace direct consultation of the primary guideline.

Acknowledgments: During the preparation of this manuscript, the authors used a generative AI tool, Claude (Anthropic), to assist with writing and language editing, formatting, and reference management. AI tools were not used for study design, data collection, statistical analysis, or figure generation, which were performed by the authors. All authors reviewed and edited the entire manuscript, including AI-assisted text, and take full responsibility for the integrity, originality, and accuracy of the content. No AI tool is listed as an author.

Funding: This research received no external funding.

Conflict of Interest: The authors declare no conflict of interest.

Informed Consent: Not applicable; the study did not involve human participants.

Ethical Approval: This study did not involve human participants, patient-level identifiers, or animal subjects; therefore, ethics committee approval was not required.

Author Contributions

Concept and design: A.Ş., H.S.G. Data collection: A.Ş., S.G. Blinded response scoring: S.G., H.S.G. Statistical analysis: H.S.G. Drafting of the manuscript: A.Ş., S.G., H.S.G. Critical revision of the manuscript: H.S.G. Supervision: H.S.G. All authors read and approved the final manuscript.

Data Availability

The data presented in this study are available from the corresponding author on reasonable request.

Supplementary Materials

The following supplementary materials are available as separate files.

Supplementary Table S1. TRIPOD-LLM reporting checklist.
Supplementary Table S2. Complete question set with gold-standard answers.
Supplementary Table S3. Full prompt texts for the three prompting strategies.
Supplementary Table S4. All 300 LLM responses with corresponding reviewer scores.
Supplementary Table S5. Three-tier scoring rubric.
Supplementary Table S6. Model estimation, prior-sensitivity, and validation diagnostics.
Supplementary Table S7. Discordant responses.
Supplementary Table S8. Concordance rates by clinical domain and prompt strategy.
Supplementary Table S9. Exploratory clustered analyses (question type and recommendation strength).
Supplementary Table S10. Domain × error-type cross-tabulation by reviewer.
Supplementary Table S11. Illustrative difficult examples.
Supplementary Table S12. Consistently non-concordant questions.
Supplementary Figure S1. Study design and workflow.
Supplementary Figure S2. Heat map of concordance rates by clinical domain and prompt strategy.

Declaration of Generative AI and AI-Assisted Technologies in the Writing Process: During the preparation of this work, the authors used a generative AI tool, Claude (Anthropic), to assist with writing and language editing, formatting, and reference management. AI tools were not used for study design, data collection, statistical analysis, or figure generation. The authors reviewed and edited all content, take full responsibility for its integrity, originality, and accuracy, and confirm that no AI tool is listed as an author.

REFERENCES

1. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med. 2023;29(8):1930–1940. https://doi.org/10.1038/s41591-023-02448-8
2. Carl N, Schramm F, Haggenmüller S, et al. Large language model use in clinical oncology. npj Precis Oncol. 2024;8(1):240. https://doi.org/10.1038/s41698-024-00733-4
3. Fu L, Tian T, Yao K, et al. Global pattern and trends in penile cancer incidence: population-based study. JMIR Public Health Surveill. 2022;8(7):e34874. https://doi.org/10.2196/34874
4. Huang J, Chan SC, Pang WS, et al. Incidence, risk factors, and temporal trends of penile cancer: a global population-based analysis. BJU Int. 2024;133(3):314–323. https://doi.org/10.1111/bju.16224
5. Visser O, Adolfsson J, Rossi S, et al. Incidence and survival of rare urogenital cancers in Europe. Eur J Cancer. 2012;48(4):456–464. https://doi.org/10.1016/j.ejca.2011.10.031
6. Cindolo L, Spiess PE, Bada M, et al. Adherence to EAU guidelines on penile cancer translates into better outcomes: a multicenter international study. World J Urol. 2019;37(8):1649–1657. https://doi.org/10.1007/s00345-018-2549-3
7. Eppler M, Ganjavi C, Ramacciotti LS, et al. Awareness and use of ChatGPT and large language models: a prospective cross-sectional global survey in urology. Eur Urol. 2024;85(2):146–153. https://doi.org/10.1016/j.eururo.2023.10.014
8. Nedbal C, Naik N, Castellani D, Gauhar V, Geraghty R, Somani BK. ChatGPT in urology practice: revolutionizing efficiency and patient care with generative artificial intelligence. Curr Opin Urol. 2024;34(2):98–104. https://doi.org/10.1097/MOU.0000000000001151
9. Tsai CY, Cheng PY, Deng JH, Jaw FS, Yii SC. ChatGPT v4 outperforming v3.5 on cancer treatment recommendations in quality, clinical guideline, and expert opinion concordance. Digit Health. 2024;10:20552076241269538. https://doi.org/10.1177/20552076241269538
10. Ozgor F, Caglar U, Halis A, et al. Urological cancers and ChatGPT: assessing the quality of information and possible risks for patients. Clin Genitourin Cancer. 2024;22(2):454–457.e4. https://doi.org/10.1016/j.clgc.2023.12.017
11. 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. https://doi.org/10.1089/end.2024.0860
12. Hao Y, Qiu Z, Holmes J, et al. Large language model integrations in cancer decision-making: a systematic review and meta-analysis. npj Digit Med. 2025;8(1):450. https://doi.org/10.1038/s41746-025-01824-7
13. Ao G, Chen M, Li J, Nie H, Zhang L, Chen Z. Comparative analysis of large language models on rare disease identification. Orphanet J Rare Dis. 2025;20(1):150. https://doi.org/10.1186/s13023-025-03656-w
14. European Association of Urology (EAU) and American Society of Clinical Oncology (ASCO). EAU-ASCO Guidelines on Penile Cancer. Arnhem: EAU Guidelines Office; 2025.
15. Gallifant J, Afshar M, Ameen S, et al. The TRIPOD-LLM reporting guideline for studies using large language models. Nat Med. 2025;31(1):60–69. https://doi.org/10.1038/s41591-024-03425-5
16. Bedi S, Liu Y, Orr-Ewing L, et al. Testing and evaluation of health care applications of large language models: a systematic review. JAMA. 2025;333(4):319–328. https://doi.org/10.1001/jama.2024.21700
17. Asgari E, Montaña-Brown N, Dubois M, et al. A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation. npj Digit Med. 2025;8(1):274. https://doi.org/10.1038/s41746-025-01670-7
18. Omar M, Sorin V, Collins JD, et al. Multi-model assurance analysis showing large language models are highly vulnerable to adversarial hallucination attacks during clinical decision support. Commun Med (Lond). 2025;5(1):330. https://doi.org/10.1038/s43856-025-01021-3

19. Cohen J. Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. Psychol Bull. 1968;70(4):213–220. https://doi.org/10.1037/h0026256

20. Jeon S, Kim HG. A comparative evaluation of chain-of-thought-based prompt engineering techniques for medical question answering. Comput Biol Med. 2025;196(Pt A):110614. https://doi.org/10.1016/j.compbiomed.2025.110614

21. Wang L, Chen X, Deng X, et al. Prompt engineering in consistency and reliability with the evidence-based guideline for LLMs. npj Digit Med. 2024;7(1):41. https://doi.org/10.1038/s41746-024-01029-4

22. Zaghir J, Naguib M, Bjelogrlic M, Névéol A, Tannier X, Lovis C. Prompt engineering paradigms for medical applications: scoping review. J Med Internet Res. 2024;26:e60501. https://doi.org/10.2196/60501

23. Chen D, Avison K, Alnassar S, Huang RS, Raman S. Medical accuracy of artificial intelligence chatbots in oncology: a scoping review. The Oncologist. 2025;30(4):oyaf038. https://doi.org/10.1093/oncolo/oyaf038

24. Zhu K, Zhang J, Klishin A, et al. Evaluating the accuracy of responses by large language models for information on disease epidemiology. Pharmacoepidemiol Drug Saf. 2025;34(2):e70111. https://doi.org/10.1002/pds.70111

25. Tung JYM, Le Q, Yao J, et al. Performance of retrieval-augmented generation large language models in guideline-concordant prostate-specific antigen testing: comparative study with junior clinicians. J Med Internet Res. 2025;27:e78393. https://doi.org/10.2196/78393