International Journal of Science and Research Archive
eISSN: 2582-8185
Cross Ref DOI: 10.30574/ijsra
Journal homepage: https://ijsra.net/

(RESEARCH ARTICLE)

Check for updates

Chatbot-generated versus urology resident-generated responses to common patient questions: A blinded evaluation by senior faculty

Theodor Florin Pantilimonescu [1,*], Denisa Oana Zelinschi [2], Alexandra Lazan [2] and Viorel Dragoș Radu [3]

[1] *Department of Physiology, "Grigore T. Popa" University of Medicine and Pharmacy, 700115 Iași, Romania.*
[2] *Department of Obstetrics and Gynaecology, "Grigore T. Popa" University of Medicine and Pharmacy, 700115 Iași, Romania.*
[3] *Department of Urology, "Grigore T. Popa" University of Medicine and Pharmacy, 700115 Iași, Romania.*

International Journal of Science and Research Archive, 2026, 19(02), 582-591

Publication history: Received on 02 April 2026; revised on 09 May 2026; accepted on 11 May 2026

Article DOI: https://doi.org/10.30574/ijsra.2026.19.2.1071

Abstract

Objective: To compare the quality of responses generated by a GPT-4.1-based chatbot, trained to respond strictly based on EAU and AUA clinical guidelines, with those produced by a collaborative panel of urology residents when answering common patient questions, as evaluated by blinded senior faculty members.

Methods: Ten clinical questions most frequently encountered in urological practice were selected by a resident panel covering pre-operative, post-operative, diagnostic, and oncological categories. Chatbot responses were translated into Romanian prior to evaluation. Three senior faculty members, blinded to response source, evaluated each response using a five-metric Likert scale (1–5): clinical accuracy, completeness, safety, empathy/language, and actionability. Mann-Whitney U tests with Bonferroni correction were used for group comparisons.

Results: The chatbot achieved significantly higher scores than residents for completeness (3.90 vs. 2.90, U = 86.5, p = 0.002, r = 0.730) and actionability (3.70 vs. 2.70, U = 86.0, p = 0.004, r = 0.720), both with large effect sizes. The overall composite score favoured the chatbot (3.66 vs. 2.86, p = 0.008). Empathy scores were comparably low in both groups. One chatbot clinical hallucination event was documented, yielding the lowest score of the series.

Conclusion: A GPT-4.1-based chatbot trained on clinical guidelines outperformed urology residents on completeness and actionability metrics. However, a clinical hallucination event highlights that mean superiority does not ensure consistent safety at the individual response level.

Keywords: Artificial Intelligence; Urology; Patient Communication; Chatbot; Clinical Safety; Blinded Evaluation

1. Introduction

Effective communication between healthcare providers and patients is a cornerstone of quality medical care. A meta-analysis by Zolnierek and DiMatteo encompassing 106 studies demonstrated that physician communication is significantly correlated with patient adherence, with a 19% higher risk of non-adherence among patients whose physicians communicate poorly [1]. In urology specifically, patients frequently face anxiety-provoking diagnoses such as prostate cancer, complex surgical decisions, and sensitive topics including sexual function and urinary continence, making clear and empathetic communication particularly critical [2].

The rapid advancement of large language models (LLMs) has introduced a new paradigm in medical communication. In a landmark cross-sectional study, Ayers et al. compared physician and chatbot responses to 195 patient questions from

* Corresponding author: Theodor Florin Pantilimonescu; ORCID: https://orcid.org/0009-0004-5860-3696

a public social media forum and found that chatbot responses were preferred by evaluators 79% of the time, rated 3.6 times higher for quality and 9.8 times higher for empathy compared to physician responses [3]. Within urology, several studies have begun to evaluate AI performance. Javid et al. prospectively compared ChatGPT responses with urologist responses to 61 kidney stone patient questions and reported comparable accuracy but variable empathy scores using a blinded expert evaluation design [4]. Beyatlı et al. evaluated ChatGPT-4 responses to 77 upper tract urothelial carcinoma questions against EAU 2025 guidelines and found an overall accuracy rate of 92.2% [5]. Şahin et al. similarly assessed ChatGPT performance across urological knowledge domains using expert evaluation [6].

However, the literature presents several important limitations. First, the majority of published comparisons have utilised general-purpose AI models without domain-specific configuration, limiting the clinical applicability of findings [2]. Second, AI-generated medical content remains susceptible to hallucinations — instances where the model produces fabricated or clinically incorrect information with apparent confidence — which pose a distinct safety risk in patient communication [7]. Third, a recent systematic review by Bhuiyan et al. examining AI tools for patient communication and shared decision-making in urology identified only 14 eligible studies and concluded that evidence remains limited, with most studies not evaluating AI tools specifically designed for clinical communication or health literacy [2].

The emergence of customisable chatbot platforms, such as Microsoft Copilot Studio, now enables clinicians and researchers to build domain-specific AI tools within institutional frameworks, using state-of-the-art LLMs without extensive programming expertise [8]. This raises the question of whether such purpose-built chatbots can meaningfully complement resident-level clinical communication in a standardised evaluation framework.

The present study aimed to compare the quality of responses generated by a custom chatbot, built using Microsoft Copilot Studio with GPT-4.1 as the underlying language model and trained to respond strictly based on EAU and AUA clinical guidelines, with responses produced by a collaborative panel of urology residents (years R2, R4, and R5), when answering the ten most frequently encountered patient questions in urological practice. Responses were evaluated by a blinded panel of three senior faculty members using a structured five-metric Likert scale encompassing clinical accuracy, completeness, safety, empathy/language, and actionability.

2. Materials and Methods

2.1. Study design

This was a single-centre, blinded, comparative evaluation study conducted at the "Grigore T. Popa" University of Medicine and Pharmacy and the "Dr. C.I. Parhon" University Clinical Hospital, Iași, Romania. The study compared responses generated by a custom AI chatbot with those produced by a collaborative panel of urology residents, as evaluated by blinded senior faculty members. The study was approved by the Ethics Committee of the "Dr. C.I. Parhon" University Clinical Hospital (approval no. 4703/15 May 2025).

2.2. Question selection

Ten clinical questions were selected by the resident panel as the most frequently encountered patient inquiries in urological practice. Questions were chosen through purposive sampling to represent a balanced distribution across four clinical categories: pre-operative (n = 4), post-operative (n = 3), diagnostic/interpretive (n = 2), and oncological/technical (n = 1). The complete list of questions is provided in Table 1. No real patient data or identifiable patient information were used; all questions were formulated to represent typical clinical inquiries based on the residents' collective clinical experience.

2.3. Response generation

Two independent groups generated responses to the same set of ten questions.

Group 1 — Resident panel (control): Three urology residents at different training stages collaborated to produce a single consensus response for each question. The panel comprised a second-year resident (R2), responsible for adapting language to patient comprehension level; a fourth-year resident (R4), responsible for ensuring diagnostic and treatment protocol accuracy per AUA 2026 guidelines; and a fifth-year resident and chief resident (R5), responsible for final supervision and alignment with EAU 2026 guidelines. The residents had full access to the current editions of EAU and AUA guideline documents during response generation. A single consolidated response per question was produced through deliberation.

Group 2 — AI chatbot (intervention): A custom chatbot was built using Microsoft Copilot Studio under the institutional license of the "Grigore T. Popa" University of Medicine and Pharmacy. The underlying large language model was GPT-4.1 (OpenAI). The chatbot was trained to respond strictly based on EAU and AUA clinical guidelines, with the goal of providing the most accurate clinical information possible. All ten questions were submitted within a single continuous conversational thread, allowing the chatbot access to previously disclosed patient information across sequential questions. Chatbot responses were generated in English and subsequently translated into Romanian prior to evaluation. The complete system prompt used for chatbot configuration is provided in Supplementary Material 1.

2.4. Blinding and evaluation procedure

Three senior faculty members from the Departments of Urology and Physiology at the "Grigore T. Popa" University of Medicine and Pharmacy served as evaluators. Evaluators were blinded to the source of each response; responses were presented in randomised order, labelled only as "Response A" and "Response B" for each question. No identifying information regarding the authorship of responses was disclosed to evaluators at any point during the evaluation process.

2.5. Evaluation instrument

Each response was assessed using a structured evaluation grid comprising five metrics, each scored on a Likert scale from 1 (lowest) to 5 (highest): (a) clinical accuracy: degree of adherence to current EAU 2026 and AUA 2026 guideline recommendations; (b) completeness: coverage of all critical clinical information, including nuances, alternative options, and levels of evidence; (c) safety: absence of clinically dangerous omissions, errors, or misleading information that could harm the patient; (d) empathy and language: appropriateness of language for the patient's level of understanding, without being condescending; and (e) actionability: whether the patient would know what concrete steps to take after reading the response.

In addition, three binary items were recorded for each response: (i) "Would you send this response to a patient without modifications?" (Yes / Yes with minor edits / Yes with major edits / No); (ii) presence of clinically dangerous omissions or red flags (Yes / No); and (iii) presence of clinical hallucination or incorrect medical context (Yes / No).

Evaluation scores were determined through a structured consensus process among the three evaluators, following independent review of each response. Each evaluator first reviewed responses within their domain of expertise: clinical accuracy and completeness were primarily assessed by the guideline expert, safety by the patient safety specialist, and empathy/language and actionability by the medical communication expert. Consensus scores were then established through panel deliberation for each metric, consistent with modified Delphi-method principles [9].

2.6. Statistical analysis

All statistical analyses were performed using Python (version 3.x) with the SciPy library. Given the ordinal nature of Likert-scale data and the small sample size (N = 10 questions per group), non-parametric tests were used throughout. Normality of score distributions was assessed using the Shapiro-Wilk test and confirmed the appropriateness of non-parametric methods.

Group comparisons for each of the five Likert metrics were performed using the Mann-Whitney U test (two-tailed, exact p-values). Effect sizes were quantified using rank-biserial correlation (r), interpreted as small (r = 0.10), medium (r = 0.30), or large (r = 0.50). Bonferroni correction was applied to account for multiple comparisons across the five metrics, yielding an adjusted significance threshold of $\alpha = 0.010$ per test. The overall composite score (mean of all five metrics per response) was compared using Mann-Whitney U at $\alpha = 0.05$ without Bonferroni correction.

Binary outcomes (clinical hallucination, red flag presence, patient-ready verdict) were compared between groups using Fisher's exact test, with odds ratios and exact p-values reported. A secondary descriptive analysis compared mean scores across question categories. Due to the small subgroup sizes ($n \leq 3$ per category), no inferential statistics were applied to subgroup comparisons.

Inter-rater reliability was not calculated as a separate metric, as evaluation scores represented the consensus of the three-member expert panel following structured deliberation, rather than independent ratings. This approach is acknowledged as a limitation in the Discussion section.

3. Results

3.1. Overview

A total of ten clinical questions were evaluated, yielding 20 responses (10 resident-generated, 10 chatbot-generated) and 100 individual Likert scores (20 responses × 5 metrics).

Table 1 Clinical questions used in the study (N = 10)

ID	Category	Patient question	Type
URO-01	Pre-operative	Doctor, someone told me that prostate surgery leaves you without your manhood. Is that true?	Emotional
URO-02	Pre-operative	What tests do I absolutely need before prostate surgery?	Procedural
URO-03	Diagnostic	My PSA came back at 12 ng/mL. Does that mean I definitely have cancer?	Interpretive
URO-04	Pre-operative	Do I need to stop aspirin before surgery? How long before?	Operational
URO-05	Diagnostic	My Gleason score is 7 (3+4). Am I in the favourable intermediate-risk group?	Interpretive
URO-06	Pre-operative	I am afraid of general anaesthesia. Can I have spinal anaesthesia instead?	Emotional
URO-07	Post-operative	How long will I stay in the hospital if I have robotic prostatectomy?	Planning
URO-08	Oncologic	What is the difference between laparoscopic and robotic in terms of positive surgical margins?	Technical
URO-09	Post-operative	Why do I have blood in my urine after surgery? Is it serious or normal?	Emotional
URO-10	Post-operative	The catheter hurts a lot. Can I take it out earlier than scheduled?	Operational

3.2. Descriptive statistics

The chatbot achieved higher mean scores than the resident panel on four of five metrics (Table 2). The largest between-group differences were observed for completeness (chatbot: 3.90 ± 0.74 vs. residents: 2.90 ± 0.32, Δ = +1.00), safety (4.30 ± 1.06 vs. 3.30 ± 0.67, Δ = +1.00), and actionability (3.70 ± 0.48 vs. 2.70 ± 0.67, Δ = +1.00). Clinical accuracy scores were higher for the chatbot but with a smaller difference (3.90 ± 0.99 vs. 3.40 ± 0.70, Δ = +0.50). Empathy and language scores were the lowest of all metrics for both groups and showed the smallest between-group difference (chatbot: 2.50 ± 0.71 vs. residents: 2.00 ± 0.47, Δ = +0.50). The overall composite mean score favoured the chatbot (3.66 ± 0.67 vs. 2.86 ± 0.45, Δ = +0.80).

Table 2 Descriptive statistics per metric per group (N = 10)

Metric	Group	Mean	SD	Median	IQR	Min	Max	Δ
Clinical Accuracy	Residents	3.40	0.70	3.5	1.0	2	4	
	Chatbot	3.90	0.99	4.0	1.5	2	5	+0.50
Completeness	Residents	2.90	0.32	3.0	0.0	2	3	
	Chatbot	3.90	0.74	4.0	1.0	3	5	+1.00
Safety	Residents	3.30	0.67	3.0	1.0	2	4	
	Chatbot	4.30	1.06	5.0	1.0	2	5	+1.00

Empathy/Language	Residents	2.00	0.47	2.0	0.0	1	3	
	Chatbot	2.50	0.71	2.0	1.0	2	4	+0.50
Actionability	Residents	2.70	0.67	3.0	1.0	2	4	
	Chatbot	3.70	0.48	4.0	1.0	3	4	+1.00
Overall composite	Residents	2.86	0.45	3.0	0.6	2.0	3.4	
	Chatbot	3.66	0.67	3.7	0.7	2.4	4.6	+0.80

3.3. Inferential statistics

Shapiro-Wilk testing confirmed non-normal distributions for 9 of 10 metric-group combinations ($p < 0.05$), supporting the use of non-parametric methods (Table 3).

Table 3 Shapiro-Wilk normality tests (N = 10 per group)

Metric	Group	W	p-value	Normal?
Clinical Accuracy	Residents	0.781	0.009	No
	Chatbot	0.886	0.152	Yes
Completeness	Residents	0.366	<0.001	No
	Chatbot	0.833	0.036	No
Safety	Residents	0.802	0.015	No
	Chatbot	0.731	0.002	No
Empathy/Language	Residents	0.658	<0.001	No
	Chatbot	0.731	0.002	No
Actionability	Residents	0.802	0.015	No
	Chatbot	0.594	<0.001	No

Mann-Whitney U tests revealed statistically significant differences between groups for two of five metrics after Bonferroni correction (Table 4). Completeness scores were significantly higher for the chatbot ($U = 86.5$, $Z = 2.759$, $p = 0.002$, Bonferroni-adjusted $p = 0.010$), with a large effect size ($r = 0.730$). Actionability scores were also significantly higher for the chatbot ($U = 86.0$, $Z = 2.721$, $p = 0.004$, Bonferroni-adjusted $p = 0.020$), with a large effect size ($r = 0.720$). Safety scores favoured the chatbot with a large effect size ($U = 80.0$, $p = 0.020$, $r = 0.600$) but did not reach statistical significance after Bonferroni correction (adjusted $p = 0.101$). Clinical accuracy ($U = 66.5$, $p = 0.197$, $r = 0.330$) and empathy/language ($U = 68.5$, $p = 0.092$, $r = 0.370$) did not reach significance, though both showed medium effect sizes favouring the chatbot. The overall composite score was significantly higher for the chatbot ($U = 85.5$, $p = 0.008$, $r = 0.710$), with a large effect size.

Table 4 Mann-Whitney U tests — Chatbot vs. Residents

Metric	U	Z	p (exact)	p (Bonf.)	r	Effect	Significant?
Clinical Accuracy	66.5	1.247	0.197	0.986	0.330	Medium	No
Completeness	86.5	2.759	0.002	0.010	0.730	Large	Yes
Safety	80.0	2.268	0.020	0.101	0.600	Large	No†
Empathy/Language	68.5	1.398	0.092	0.458	0.370	Medium	No
Actionability	86.0	2.721	0.004	0.020	0.720	Large	Yes
Overall composite	*85.5*	*2.684*	*0.008*	—	*0.710*	*Large*	*Yes*‡

Bonferroni correction applied across 5 metrics: $\alpha = 0.050/5 = 0.010$. † Significant uncorrected ($p = 0.020$) but not after Bonferroni correction. ‡ Overall composite tested at $\alpha = 0.05$ without Bonferroni correction. r = rank-biserial correlation.

3.4. Binary outcomes

Fisher's exact tests revealed a statistically significant difference for language correctness only (Table 5). The resident panel produced all responses in Romanian (10/10, 100%), whereas the chatbot generated responses in the correct language in only 1 of 10 cases (10%; $p < 0.001$). One clinical hallucination event was documented in the chatbot group (1/10, 10%) and none in the resident group (0/10, 0%); this difference was not statistically significant ($p = 1.000$). In case URO-10, the chatbot introduced the clinical context of priapism and shunt procedures in response to a question about catheter-related pain following a standard urological procedure. The chatbot produced the only three responses rated "Yes, without changes" by the expert panel (URO-04, URO-05, URO-08; 30%), whereas no resident response received this verdict (0%; $p = 0.211$). The proportion of responses rated as acceptable in any form was comparable between groups (residents: 8/10, 80%; chatbot: 7/10, 70%; $p = 1.000$).

Table 5 Binary outcomes — Fisher's exact tests

Outcome	Residents	Chatbot	p-value
Correct language (Romanian)	10/10 (100%)	1/10 (10%)	<0.001
Clinical hallucination	0/10 (0%)	1/10 (10%)	1.000
Typo / quality error	0/10 (0%)	1/10 (10%)	1.000
Verdict: Yes, without changes	0/10 (0%)	3/10 (30%)	0.211
Verdict: Yes, with minor edits	4/10 (40%)	2/10 (20%)	
Verdict: Yes, with major edits	4/10 (40%)	2/10 (20%)	
Verdict: No	2/10 (20%)	3/10 (30%)	
Verdict: Any Yes form	8/10 (80%)	7/10 (70%)	1.000

3.5. Secondary analysis — performance by question type

Descriptive comparison of mean composite scores across question categories revealed a differential pattern of performance (Table 6). The chatbot demonstrated the largest advantage on interpretive, guideline-intensive questions (chatbot mean: 4.33, residents mean: 2.67, $\Delta = +1.67$). For operational/practical planning questions, the resident panel marginally outperformed the chatbot (residents mean: 3.20, chatbot mean: 3.13, $\Delta = -0.07$). This result was driven primarily by the chatbot hallucination event in URO-10. These subgroup results are presented as descriptive trends only and should be interpreted with caution given the small subgroup sizes ($n \leq 3$ per category).

Table 6 Mean composite scores by question type (descriptive only)

Question type	Cases	Residents	Chatbot	Δ
Interpretive / guideline-intensive	URO-03, 05, 08	2.67	4.33	+1.67
Emotional / anxiety-laden	URO-01, 06, 09	2.87	3.80	+0.93
Procedural	URO-02	2.40	2.80	+0.40
Operational / practical	URO-04, 07, 10	3.20	3.13	−0.07

Subgroup N ≤ 3 per category. No inferential statistics applied.

4. Discussion

4.1. Principal findings

This pilot study compared the quality of responses generated by a GPT-4.1-based chatbot, trained to respond strictly based on EAU and AUA clinical guidelines, with those produced by a collaborative panel of urology residents, as evaluated by blinded senior faculty members. The chatbot demonstrated statistically significant superiority on two of five metrics after Bonferroni correction — completeness ($p = 0.002$, $r = 0.730$) and actionability ($p = 0.004$, $r = 0.720$) — both with large effect sizes. The overall composite score significantly favoured the chatbot ($p = 0.008$, $r = 0.710$). Safety scores also favoured the chatbot with a large effect size ($r = 0.600$) but did not survive correction for multiple

comparisons. Empathy and language scores were comparably low in both groups, representing the weakest performance dimension across the entire study. Critically, one clinical hallucination event was documented in the chatbot group, introducing an entirely irrelevant medical context and yielding the lowest score of the series.

4.2. Comparison with existing literature

Our findings are broadly consistent with the growing body of evidence demonstrating that AI-generated clinical responses can match or exceed physician-generated content on structured evaluation metrics. Ayers et al. reported that chatbot responses were preferred over physician responses 79% of the time [3]. Within urology, Javid et al. found comparable accuracy between ChatGPT and urologists in a kidney stone counselling study [4], while Beyatlı et al. reported 92.2% overall accuracy for ChatGPT-4 on upper tract urothelial carcinoma questions [5].

Recent evidence from other specialties reinforces these findings. In a comparative study of AI models and human experts in dyslipidemia management, AI achieved a 91% correct response rate compared to 72% for professors, 50% for specialists, and 21–32% for residents, with AI excelling particularly in literal guideline application [13]. Similarly, the PERFORM study, which compared eight AI models against obstetrics-gynaecology residents across 60 clinical scenarios, reported superior overall AI accuracy (73.75% vs. 65.35%) and maximum performance enhancement when AI augmented early-career residents (+29.7%) [14].

Our study extends these findings in several respects. First, our chatbot was purpose-built using Microsoft Copilot Studio and trained to respond strictly based on clinical guidelines, representing a more realistic implementation scenario [8]. Context-specific chatbots configured with relevant clinical guidelines have been shown to outperform generic AI models in guideline adherence tasks [15]. Second, our control group comprised a structured collaborative panel of residents with full guideline access, making the comparison more demanding for the chatbot. Third, the continuous conversational thread design allowed the chatbot to demonstrate longitudinal clinical reasoning — a capability not evaluated in prior studies.

4.3. Guideline adherence and completeness

The most statistically robust finding was the chatbot's superiority on completeness ($\Delta = +1.00$, $p = 0.002$). This is consistent with the observation that LLMs trained on clinical guidelines can systematically retrieve relevant information that human clinicians may omit due to memory limitations or cognitive load [13]. Resident completeness scores exhibited near-zero variance (SD = 0.32, with 8 of 10 cases scoring exactly 3/5), suggesting adequate but incomplete responses. The chatbot showed greater variability (SD = 0.74) but consistently higher performance.

Notably, the reliability of AI-generated medical information against clinical guidelines has shown mixed results. Whereas some studies have reported high concordance rates [5,13], others have demonstrated that AI chatbot recommendations frequently do not align with clinical practice guidelines, with one to two-thirds of recommendations potentially inappropriate depending on the model and domain [15]. This underscores the importance of configuring chatbots with domain-specific guideline access rather than relying on general-purpose models.

4.4. The empathy gap

Perhaps the most clinically significant finding was the consistently low empathy and language scores in both groups (residents: 2.00 ± 0.47; chatbot: 2.50 ± 0.71). Across all ten cases, neither group produced a response containing explicit emotional validation of the patient's expressed fears or concerns. This aligns with documented evidence of empathy decline during medical residency training. Noordman et al. reported a downward trend in residents' empathic communication skills and demonstrated that structured programs can improve empathy scores [10]. Byrne et al. emphasised the need for longitudinal, experiential approaches to empathy development in medical curricula [11]. The fact that the chatbot performed comparably to residents on this metric suggests that the empathy deficit may be primarily a deficit of communication structure rather than emotional experience.

4.5. Health literacy and actionability

The chatbot's significant advantage on actionability ($\Delta = +1.00$, $p = 0.004$) has important implications for patient health literacy. Swisher et al. showed that LLM-revised patient handouts achieved significantly better readability scores, with a reduction from university-level to seventh-grade reading level [16]. A scoping review by Aydin et al. encompassing 201 studies identified six key themes in LLM patient education applications, though the authors cautioned that varying levels of medical knowledge among patients may impede their ability to critically assess LLM-generated information [17]. In our study, the chatbot consistently structured responses with clear next steps, specific questions for the patient

to ask their clinician, and offers to personalise recommendations — elements systematically absent from resident responses.

4.6. The hallucination event

The clinical hallucination documented in URO-10, where the chatbot introduced the irrelevant context of priapism and shunt procedures in response to a catheter-related pain question, represents a qualitatively distinct risk category absent from resident-generated responses. While residents produced responses with omissions, no resident response contained fabricated clinical context. Bélisle-Pipon argued that LLMs are trained to predict patterns rather than to verify clinical accuracy [7]. Walker et al. reported wrong or out-of-context answers when ChatGPT was evaluated against clinical guidelines [18]. In our study, the hallucination was presented with the same confidence as correct responses, offering no signal that the information was fabricated. This represents a fundamental limitation of current LLM architecture.

The hallucination event had a measurable impact on results, producing the only case in which residents outperformed the chatbot and shifting the operational question subgroup from a chatbot advantage to a marginal resident advantage.

4.7. Complementary roles

The secondary analysis by question type revealed that the chatbot demonstrated its largest advantage on interpretive, guideline-intensive questions ($\Delta = +1.67$), while residents performed comparably on operational questions ($\Delta = -0.07$). This suggests complementary rather than substitutive roles, consistent with Altamimi et al. who argued that AI chatbots should function as supplements to medical professionals [12]. The PERFORM study further supports this, demonstrating maximum performance enhancement when AI augmented early-career residents [14].

4.8. Limitations

This study has several limitations. First, the sample size of ten questions limits statistical power and precludes definitive conclusions. Second, consensus scoring precluded calculation of inter-rater reliability. Third, the continuous conversational thread granted the chatbot access to cumulative patient information not available to residents. Fourth, chatbot responses were translated from English, which may have affected empathy ratings. Fifth, generalisability is limited to the specific LLM version (GPT-4.1), platform (Microsoft Copilot Studio), and clinical domain (urology). Finally, as a single-centre study, institutional biases cannot be excluded.

5. Conclusion

This blinded pilot study demonstrated that a GPT-4.1-based chatbot, configured to respond strictly based on EAU and AUA clinical guidelines, produced responses to common urological patient questions that were rated significantly higher than those generated by a collaborative panel of urology residents on completeness and actionability metrics, with large effect sizes. The overall composite score significantly favoured the chatbot. Both groups performed poorly on empathy and emotional validation, suggesting a systematic communication gap that transcends the human-AI distinction. Importantly, a single clinical hallucination event demonstrated that mean superiority in AI performance does not guarantee consistent safety at the individual response level. These findings support a complementary model of AI integration in urological patient communication, where chatbot-generated draft responses are reviewed and personalised by clinicians. A larger-scale study with an expanded question set and independent evaluator scoring is warranted to confirm these preliminary findings.

Compliance with ethical standards

Acknowledgments

The authors wish to thank the staff of the “Dr. C.I. Parhon” University Clinical Hospital and the “Grigore T. Popa” University of Medicine and Pharmacy, Iaşi, for their institutional support throughout this study.

Disclosure of conflict of interest

The authors declare no conflict of interest.

Statement of ethical approval

This study was approved by the Ethics Committee of the "Dr. C.I. Parhon" University Clinical Hospital (approval no. 4703/15 May 2025).

Statement of informed consent

All participants provided informed consent prior to study participation. The study was conducted in accordance with the principles of the Declaration of Helsinki.

References

[1] Zolnierek KB, DiMatteo MR. Physician communication and patient adherence to treatment: a meta-analysis. Med Care. 2009;47(8):826–34.

[2] Bhuiyan N, Bracey S, Pietropaolo A, Jones P, Tzelves L, Somani BK. The role of artificial intelligence in improving patient communication and shared decision-making in urology: A systematic review. Scottish medical journal [Internet]. 2026 May;369330261418599. Available from: https://pubmed.ncbi.nlm.nih.gov/41642909/.

[3] Ayers JW, Poliak A, Dredze M, Lennon EC, Zhu Z, Kelley D, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med. 2023;183(6):589–96.

[4] Javid M, Bhandari M, Parameshwari P, Reddiboina M, Prasad S. Evaluation of ChatGPT for patient counseling in kidney stone clinic: a prospective study. J Endourol. 2024;38(10):1063–70.

[5] Beyatlı M, Güngör HS, İnkaya A, Sobay R, Tahra A, Küçük EV. Expert evaluation of ChatGPT-4 responses to upper tract urothelial carcinoma questions: a prospective comparative study. J Clin Med. 2025;14(18):6353.

[6] Şahin B, Genç YE, Doğan K, Şener TE, Şekerci ÇA, Tanıdır Y, et al. Evaluating the performance of ChatGPT in urology: a comparative study of knowledge interpretation and patient guidance. J Endourol. 2024;38(2):186–93.

[7] Bélisle-Pipon JC. Why we need to be careful with LLMs in medicine. Front Med. 2024;11:1495582.

[8] Microsoft. Healthcare Agent Service in Copilot Studio. Microsoft Cloud for Healthcare; 2025. Available from: https://techcommunity.microsoft.com/blog/healthcareandlifesciencesblog/highlights-from-ignite-2025-how-agentic-ai-and-microsoft-copilot-are-empowering-/4474658.

[9] Nasa P, Jain R, Juneja D. Delphi methodology in healthcare research: how to decide its appropriateness. World J Methodol. 2021;11(4):116–29.

[10] Noordman J, Post B, van Dartel AAM, Slits JMA, olde Hartman TC. Training residents in patient-centred communication and empathy: evaluation from patients, observers and residents. BMC Med Educ. 2019;19(1):128.

[11] Byrne M, Campos C, Daly S, Lok B, Miles A. The current state of empathy, compassion and person-centred communication training in healthcare: an umbrella review. Patient Educ Couns. 2024;119:108063.

[12] Altamimi I, Altamimi A, Alhumimidi AS, Altamimi A, Temsah MH. Artificial intelligence (AI) chatbots in medicine: a supplement, not a substitute. Cureus. 2023;15(6):e40922.

[13] Ucdal M, Yurtsever K, Yildiz P, Akalin A, Mert KU, Guven GS. Comparison of Artificial Intelligence Models and Human Experts in Managing Dyslipidemia: Assessment of Adherence to Clinical Guidelines. Cureus. 2025 Aug 31.

[14] Martinelli C, Giordano A, Carnevale V, Burk SR, Porto L, Vizzielli G, et al. The PERFORM Study: Artificial Intelligence Versus Human Residents in Cross-Sectional Obstetrics-Gynecology Scenarios Across Languages and Time Constraints. Mayo Clinic Proceedings: Digital Health [Internet]. 2025 Mar 8;3(2):100206. Available from: https://www.sciencedirect.com/science/article/pii/S2949761225000136.

[15] Rossettini G, Bargeri S, Cook C, Guida S, Palese A, Rodeghiero L, et al. Accuracy of ChatGPT-3.5, ChatGPT-4o, Copilot, Gemini, Claude, and Perplexity in advising on lumbosacral radicular pain against clinical practice guidelines: cross-sectional study. Frontiers in Digital Health. 2025 Jun 27;7.

[16] Swisher CB, Kunjur AN, Kuan EC, Suh JD, Tajudeen BA. Enhancing health literacy: evaluating the readability of patient handouts revised by ChatGPT's large language model. Otolaryngol Head Neck Surg. 2024;171(4):1181–5.

[17] Aydin S, Mert Karabacak, Vlachos V, Margetis K. Large language models in patient education: a scoping review of applications in medicine. Frontiers in Medicine. 2024 Oct 29;11.

[18] Walker HL, Ghani S, Kuemmerli C, Nebiker CA, Müller BP, Raptis DA, et al. Reliability of medical information provided by ChatGPT: assessment against clinical guidelines and patient information quality instrument. BMJ Surg Interv Health Technol. 2023;5(1):e000163.