Supplementary Information

For the paper titled “Research through Evaluation (RtE) for Large Language Model in Patient-Clinician Communications”

1. In-Basket Bot Details

Subject demographic information retrieved from the EHR system included sex, age, race, ethnicity, preferred language, and the attending physician's name. Information collected from Aria included demographic information (sex, age, race, ethnicity, preferred language, and the attending physician's name), prostate cancer treatment-specific information (course description, plan intent, treatment orientation, radiation type, radiation oncology machine type, number of fractions, dose prescription, dose delivered, radiation technique, and treatment duration), and diagnosis details (cancer stage, ICD (International Classification of Diseases) diagnosis code and code type, onset date). Information collected from Epic included clinical notes, ordered by date. For in-basket bot, the information retrieval order starts with patient demographic details, followed by treatment details, diagnosis details, and lastly, clinical notes.

We used this final prompt for in-basket bot to generate response: "*Patient #ID has sent an in-basket message. Please generate a response to their message. Before generating the response, first retrieve the patient details, patient treatment details, patient diagnosis details, and patient clinical notes. Do not pull the in-basket messages. Retrieve all types of patient data simultaneously. In writing your response, feel free to make recommendations as if you were the attending physician (since your response will be approved by the attending physician). When handling prescriptions, prioritize over-the-counter if appropriate. Sign off as the attending physician. Do not mention that the patient should contact their provider, since you are acting as their provider. Prior to giving your response, explain your reasoning step by step in an analysis section. As part of your analysis, indicate whether the patient has provided enough information to adequately respond to the message. If you determine that the patient has not provided enough information, please ask for more information in your message to the patient. Assume the patient has a high school education level. Here is the in-basket message that you should respond to: Message Details.*"

1.2. In-Basket Messages Grading Rubric

0 - Not applicable; 1 - Strongly disagree; 2 - Disagree; 3 - Neutral; 4 - Agree; 5 - Strongly agree

Completeness (6-point scale)

Definition: The extent to which the response addresses all parts of the patient's message.

Key Points:

- Does the response cover all the questions or concerns raised by the patient?
- Are there any missing elements that the patient might need to know?
- Does the response provide a thorough and comprehensive answer?

Correctness (6-point scale)

Definition: The accuracy and reliability of the information provided in the response.

Key Points:

- Is the information factually correct?
- Are medical terms and treatment options accurately described?
- Does the response avoid any misleading or incorrect statements?

Clarity (6-point scale)

Definition: The ease with which the patient can understand the response.

Key Points:

- How clear and easy is the language to understand?
- Are medical terms explained in a way that a layperson can comprehend?
- Is the response well-organized and logically structured?

Empathy (6-point scale)

Definition: The degree to which the response shows understanding and sensitivity to the patient's emotional state.

Key Points:

- Does the response acknowledge the patient's feelings and concerns?
- Is the tone compassionate and supportive?
- Does the response make the patient feel heard and cared for?

Extensive Editing (4-point scale)

Definition: The degree to which the response can be sent to the patient directly.

Scale:

- Would use this without editing.
- Minor (<1 minute) editing.
- Major editing.
- Would not use this.

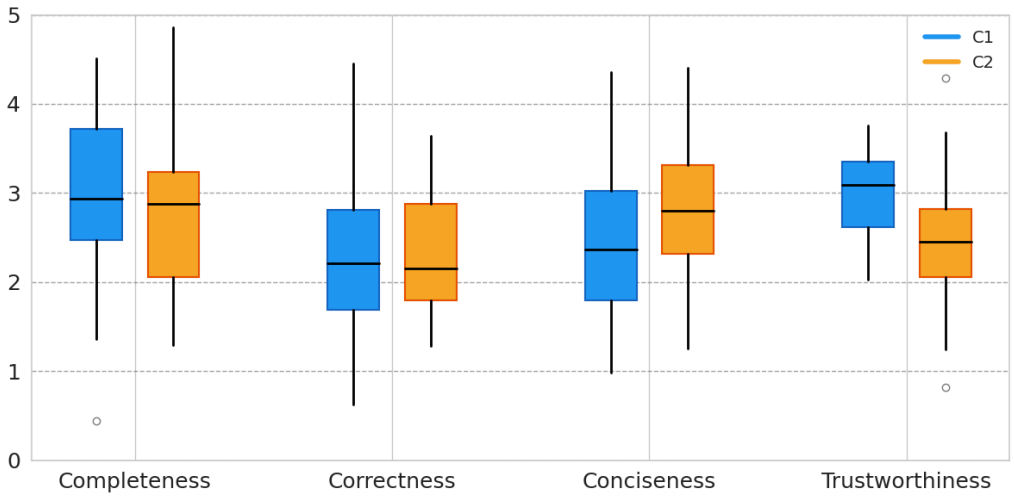
Clinicians played a key role in designing and developing the chatbot's inputs and outputs. However, the chatbot has not yet been deployed for clinical use and requires thorough evaluation, particularly by expert evaluators.

2. Detailed Grading Results

Supplementary Table 1: Direct Comparison Study (Round 1) Inter-rater reliability and mean variability metrics. ICC stands for Intraclass Correlation Coefficient, CI stands for Confidence Interval.

Metric	Completeness	Correctness	Conciseness	Trustworthiness
C1 Avg (Std)	2.59 (0.84)	2.56 (0.97)	2.59 (0.93)	2.74 (0.53)
C2 Avg (Std)	2.67 (0.92)	2.33 (0.68)	2.85 (0.72)	2.48 (0.80)
Single Raters ICC1	0.64	0.50	0.59	0.53
Single Raters CI	[0.3, 0.84]	[0.1, 0.76]	[0.23, 0.81]	[0.15, 0.78]
Average Raters ICC3k	0.77	0.71	0.79	0.76
Average Raters CI	[0.43 0.91]	[0.28 0.88]	[0.47 0.91]	[0.4 0.9]
p-value (Two Clinician Graders)	<0.001	<0.05	<0.001	<0.05

Two Graders' Krippendorff's Alpha	0.36	0.30	0.61	0.35

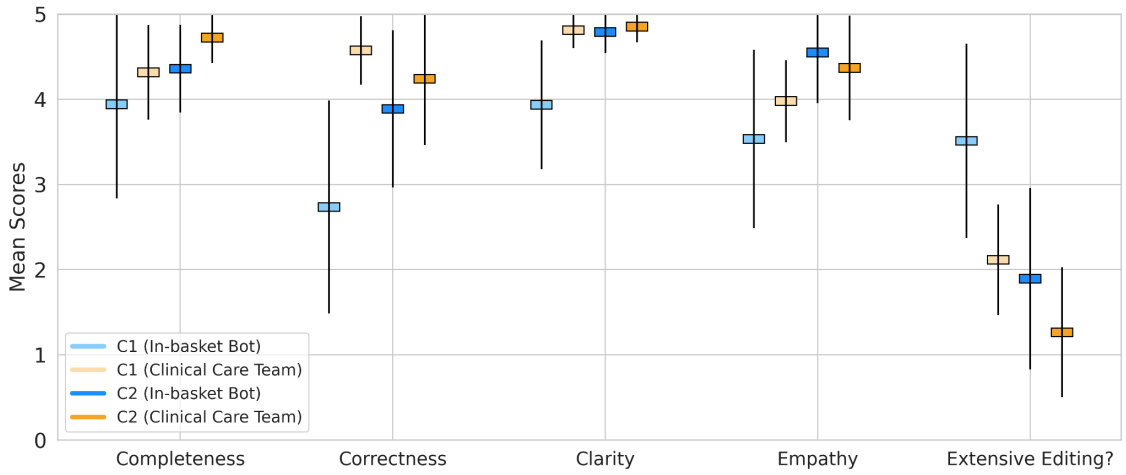


Supplementary Figure 1. Direct Comparison Study (Round 1) Evaluation Results. A score of 0 indicates not applicable, lower scores indicate a strong preference for the clinical care team's response, while higher scores indicate a preference for the in-basket bot's response.

Supplementary Table 2: Single-Blind Study (Round 2) Inter-rater reliability and mean variability metrics. ICC stands for Intraclass Correlation Coefficient, CI stands for Confidence Interval.

Metric	Completeness	Correctness	Clarity	Empathy	Extensive Editing Required
C1 In-basket Bot Avg (Std)	3.60 (1.26)	3.20 (1.48)	4.10 (0.88)	3.00 (1.05)	3.50 (1.51)
C1 Clinical Care Team Avg (Std)	4.47 (0.77)	4.68 (0.48)	4.89 (0.32)	4.16 (0.60)	2.00 (0.58)
C2 In-basket Bot Avg (Std)	4.70 (0.95)	4.30 (1.06)	4.90 (0.32)	4.70 (0.95)	2.00 (1.49)
C2 Clinical Care Team	4.84 (0.37)	4.47 (0.96)	4.95 (0.23)	4.53 (0.70)	1.53 (1.07)

Avg (Std)					
Single Raters ICC1	0.20	0.69	-0.16	-0.02	0.60
Single Raters CI	[-0.17 0.52]	[0.44 0.84]	[-0.49 0.21]	[-0.37 0.35]	[0.32 0.79]
Average Raters ICC3k	0.51	0.83	-0.23	0.32	0.88
Average Raters CI	[-0.04 0.77]	[0.63 0.92]	[-1.62 0.42]	[-0.44 0.68]	[0.74 0.94]
p-value (Two Clinician Graders)	0.03	<0.001	0.71	0.15	<0.001
Two Graders' Krippendorff's Alpha	0.19	0.69	-0.16	-0.02	0.60



Supplementary Figure 2. Single-Blind Study (Round 2) Evaluation Results. Each individual response is evaluated, with clinicians blind to whether the response originated from a human staff member or the bot. Responses are scored on a scale from 0 to 5, where 1 indicates "strongly disagree," 5 indicates "strongly agree," and 0 represents "not applicable." In the "Extensive Editing" category, lower scores reflect less editing required for the responses, making lower scores preferable for this measure.

Supplementary Table 3: Multi-Clinician Evaluation and LLM-as-a-Grader Study (Round 3 & 4) Inter-rater reliability and mean variability metrics. ICC(1) represents single raters and

ICC(3k) represents average raters. The associated p-values were obtained from F-statistics produced in the ICC analysis, testing the null hypothesis that the true ICC equals zero. For Krippendorff's alpha, reliability was calculated using a bootstrap procedure to estimate confidence intervals (CI), though no significance test was performed. All p-values reflect the probability of observing the obtained agreement level if there were truly no agreement among raters. For Krippendorff's Alpha, all values are below the 0.667 threshold typically considered the minimum for acceptable reliability.

Grader Type	**Metric**	**Completeness**	**Correctness**	**Clarity**	**Empathy**
Human Graders	Single Raters ICC1	0.21	0.17	0.04	0.22
	Single Raters CI	[0.14, 0.29]	[0.10, 0.24]	[-0.03, 0.11]	[0.15, 0.29]
	Average Raters ICC3k	0.48	0.39	0.17	0.49
	Average Raters CI	[0.37, 0.57]	[0.26, 0.50]	[-0.01, 0.32]	[0.38, 0.58]
	p-value (Three Clinician Graders)	<0.05	0.14	<0.05	0.07
	Senior Graders' Mean Variability	0.94	0.98	0.74	0.74
	Three Graders' Krippendorff's Alpha	0.63	0.81	0.75	0.53
LLM Graders	Single Raters ICC1	0.26	0.21	0.21	0.23
	Single Raters CI	[0.22, 0.31]	[0.16, 0.25]	[0.17, 0.26]	[0.18, 0.27]
	Average Raters ICC3k	0.82	0.67	0.71	0.82
	Average Raters CI	[0.79, 0.85]	[0.61, 0.72]	[0.65, 0.75]	[0.78, 0.84]
	p-value (GPT-4 Grader)	<0.001	<0.05	<0.05	<0.001

	p-value (GPT-4o Grader)	<0.001	<0.05	<0.001	<0.001
	p-value (Gemini Grader)	<0.001	<0.001	<0.001	<0.001
	p-value (Llama Grader)	<0.001	<0.001	<0.001	<0.001
	LLM Graders' Krippendorff's Alpha	0.18	0.10	0.07	0.17

Supplementary Table 4: RtE Four Rounds Evaluation Results. Descriptive statistics comparing the average performance of the In-basket Bot and Clinical Care Team across C1-C3 graders. Evaluation also includes comparisons of four LLM graders (GPT-4o, Gemini, GPT-4, and Llama 3.1) for both the In-basket Bot and Clinical Care Team on 158 pairs of in-basket message evaluations. Standard deviations are shown in parentheses.

Statistic	**Completeness**	**Correctness**	**Clarity**	**Empathy**
Clinician C1-C3: In-Basket Bot				
C1 Grader	4.44 (0.96)	4.29 (1.10)	4.88 (0.51)	4.55 (0.72)
C2 Grader	4.69 (0.73)	4.36 (0.99)	4.93 (0.34)	4.86 (0.37)
C3 Grader	4.31 (0.86)	4.23 (0.93)	4.62 (0.72)	4.43 (0.64)
Clinician C1-C3: Clinical Care Team				
C1 Grader	4.53 (0.70)	4.64 (0.70)	4.90 (0.45)	4.44 (0.81)
C2 Grader	4.77 (0.63)	4.74 (0.63)	4.90 (0.41)	4.61 (0.70)
C3 Grader	4.47 (0.83)	4.40 (0.82)	4.75 (0.59)	4.43 (0.65)
Round 3: In-Basket Bot (LLM Graders)				
Clinician Graders (Average)	4.48	4.29	4.81	4.61
GPT-4o (Zero Shot)	3.89 (0.84)	4.89 (0.41)	4.82 (0.43)	4.11 (0.98)
Gemini 1.5-Pro (Zero Shot)	3.23 (0.89)	4.23 (0.70)	4.31 (0.65)	3.28 (0.90)

GPT-4 (Zero Shot)	3.99 (0.71)	4.81 (0.22)	4.88 (0.18)	4.02 (0.53)
Llama 3.1 (Zero Shot)	3.51 (0.78)	4.62 (0.63)	4.34 (0.60)	3.03 (0.85)
Round 3: Clinical Care Team (LLM Graders)				
Clinician Graders (Average)	4.59	4.59	4.85	4.49
GPT-4o (Zero Shot)	3.38 (0.92)	4.72 (0.59)	4.49 (0.74)	3.24 (1.19)
Gemini 1.5-Pro (Zero Shot)	2.78 (0.88)	3.77 (0.88)	3.81 (0.87)	2.78 (0.97)
GPT-4 (Zero Shot)	4.68 (1.09)	4.96 (0.63)	4.97 (0.39)	4.84 (1.23)
Llama 3.1 (Zero Shot)	3.03 (0.80)	4.25 (0.68)	3.87 (0.76)	2.42 (0.94)

Supplementary Table 5: Sample Outputs from Clinical Care Team and In-Basket Bot Responses

Patient Inquiry	Clinical Team Response	In-Basket Bot Response
Clinical Team Better than In-Basket Bot		
After the ADT injection I have been getting up at night with a weak urination stream last night after first radiation treatment the stream was reduced to a dribble on two occasions. It seems OK during the day but night seems to be a challenge. Any thoughts on this issue? [Name]	Good morning Mr. [Name], We wouldn't expect your ADT shot to affect your urination stream, however that is a common side effect from radiation treatment. Do you feel like you are completely emptying your bladder when you urinate? Are you experiencing increased urinary frequency and/or urgency? We have a few options for you, just trying to get more of an idea of your symptoms. I'll wait to hear from you and then we will come up with a plan to manage the symptoms you're experiencing. [Name]	Dear Mr. [Name], Thank you for reaching out regarding your recent changes in urination. It is not uncommon to experience changes in urination patterns during androgen deprivation therapy (ADT) and radiation treatment. Here are a few suggestions that might help: Hydration: Ensure you are drinking enough water throughout the day but try to limit fluid intake a few hours before bedtime to reduce nighttime urination. Bladder Training: Try to urinate at regular intervals during the day to help train your bladder. Over-the-Counter Medications: You may consider using over-the-counter medications like phenazopyridine (Azo) to help with urinary discomfort. Please follow the instructions on the packaging. Monitor Symptoms: Keep a log of your urination patterns and any other symptoms you experience. This information can be helpful for your next appointment. If your symptoms persist or worsen, please let us know so we can further evaluate and provide additional support. Best regards, Dr. [Name]

	AvgCompleteness: 5.0 AvgCorrectness: 5.0 AvgClarity: 5.0 AvgEmpathy: 5.0 Editing Level: Without editing	AvgCompleteness: 2.3 AvgCorrectness: 2.3 AvgClarity: 5.0 AvgEmpathy: 3.3 Editing Level: Major editing
In-Basket Bot Better than Clinical Team		
I just completed cataract surgeries in [Month]. Knowing that I was having my cataract surgeries, I was advised to not take Tamsulosin which was prescribed 9-22 after prostate radiation. Would I still benefit from using this drug in that I still wake up for bathroom breaks at night? There are 90 pills, how long should I take, and when should I stop use? [Name]	Hi Mr. [Name], If you were not on tamsulosin prior to radiation treatment, you could try weaning off it. You would take it every other day for two weeks, and if your urinary symptoms (urgency, frequency, burning with urination, hesitancy with urination) don't worsen, you could stop taking it. If those symptoms worsen, you should restart taking it daily. Let us know if you have other questions. Thanks, [Name]	Dear Mr. [Name], Thank you for your message. Given your recent cataract surgeries and the history of urinary symptoms post-radiation therapy, it is reasonable to resume Tamsulosin to manage nocturia. Please take one tablet every evening and monitor your symptoms. If you notice any persistent or worsening symptoms, please let us know. Best regards, Dr. [Name]
	AvgCompleteness: 3.0 AvgCorrectness: 4.7 AvgClarity: 5.0 AvgEmpathy: 4.5 Editing Level: Would not use this	AvgCompleteness: 4.7 AvgCorrectness: 4.7 AvgClarity: 4.7 AvgEmpathy: 4.7 Editing Level: Minor (<1 minute) editing