Original Article
Prostate cancer

pISSN: 2287-4208 / eISSN: 2287-4690
World J Mens Health 2025 Oct 43(4): 918-933
https://doi.org/10.5534/wjmh.240173

The World Journal of MEN's HEALTH

The In-depth Comparative Analysis of Four Large Language AI Models for Risk Assessment and Information Retrieval from Multi-Modality Prostate Cancer Work-up Reports

Lun-Hsiang Yuan[1,2,*], Shi-Wei Huang[2,3,*], Dean Chou[1,4,5,6], Chung-You Tsai[7,8]

[1]*Department of Biomedical Engineering, National Cheng-Kung University, Tainan, Taiwan,* [2]*Department of Urology, National Taiwan University Hospital, Yunlin Branch, Yunlin, Taiwan,* [3]*Department of Urology, National Taiwan University Hospital, Taipei, Taiwan,* [4]*Miin Wu School of Computing, National Cheng-Kung University, Tainan, Taiwan,* [5]*Academy of Innovative Semiconductor and Sustainable Manufacturing, National Cheng-Kung University, Tainan, Taiwan,* [6]*National Center for High-Performance Computing, Hsinchu, Taiwan,* [7]*Division of Urology, Department of Surgery, Far Eastern Memorial Hospital, New Taipei, Taiwan,* [8]*Department of Electrical Engineering, Yuan Ze University, Taoyuan, Taiwan*

Purpose: Information retrieval (IR) and risk assessment (RA) from multi-modality imaging and pathology reports are critical to prostate cancer (PC) treatment. This study aims to evaluate the performance of four general-purpose large language model (LLMs) in IR and RA tasks.

Materials and Methods: We conducted a study using simulated text reports from computed tomography, magnetic resonance imaging, bone scans, and biopsy pathology on stage IV PC patients. We assessed four LLMs (ChatGPT-4-turbo, Claude-3-opus, Gemini-Pro-1.0, ChatGPT-3.5-turbo) on three RA tasks (LATITUDE, CHAARTED, TwNHI) and seven IR tasks. It included TNM staging, and the detection and quantification of bone and visceral metastases, providing a broad evaluation of their capabilities in handling diverse clinical data. We queried LLMs with multi-modality reports using zero-shot chain-of-thought prompting via application programming interface. With three adjudicators' consensus as the gold standard, these models' performances were assessed through repeated single-round queries and ensemble voting methods, using 6 outcome metrics.

Results: Among 350 stage IV PC patients with simulated reports, 115 (32.9%), 128 (36.6%), and 94 (26.9%) belonged to LATITUDE, CHAARTED, and TwNHI high-risk, respectively. Ensemble voting, based on three repeated single-round queries, consistently enhances accuracy with a higher likelihood of achieving non-inferior results compared to a single query. Four models showed minimal differences in IR tasks with high accuracy (87.4%–94.2%) and consistency (ICC>0.8) in TNM staging. However, there were significant differences in RA performance, with the ranking as follows: ChatGPT-4-turbo, Claude-3-opus, Gemini-Pro-1.0, and ChatGPT-3.5-turbo, respectively. ChatGPT-4-turbo achieved the highest accuracy (90.1%, 90.7%,91.6%), and consistency (ICC 0.86, 0.93, 0.76) across 3 RA tasks.

Conclusions: ChatGPT-4-turbo demonstrated satisfactory accuracy and outcomes in RA and IR for stage IV PC, suggesting its potential for clinical decision support. However, the risks of misinterpretation impacting decision-making cannot be over-

Received: Jul 11, 2024 **Revised:** Sep 1, 2024 **Accepted:** Oct 4, 2024 **Published online** Jan 1, 2025
Correspondence to: Chung-You Tsai https://orcid.org/0000-0003-0527-1830
Division of Urology, Department of Surgery, Far Eastern Memorial Hospital, No 21, Section 2, Nanya S Rd, Banqiao District, New Taipei 220, Taiwan.
Tel: +886-2-8966-700, **Fax:** +886-2-8966-5567, **E-mail:** pgtsai@gmail.com
Correspondence to: Dean Chou https://orcid.org/0000-0001-9533-1918
Department of Biomedical Engineering, National Cheng-Kung University, No.1, University Road, Tainan City 701, Taiwan.
Tel: +886-6-2757575-63417, **Fax:** +886-6-2343270, **E-mail:** dean@gs.ncku.edu.tw
*These authors contributed equally to this work as co-first authors.

looked. Further research is necessary to validate these findings in other cancers.

Keywords: ChatGPT; Decision support systems, clinical; Information storage and retrieval; Large language model; Prostatic neoplasms; Risk assessment

INTRODUCTION

Prostate cancer (PC) ranks as the second most common cancer among men globally, accounting for an estimated 1.4 million new cases and 375,000 deaths worldwide in 2020 [1]. Notably, 15% to 30% of PC patients are diagnosed with advanced disease [2]. Among metastatic PC patients, a subgroup classified as high-risk or high-volume disease would experience rapid progression and have a dismal prognosis, with a median radiological progression-free survival of 14.8 months and overall survival of 32.2 months [3]. These patients need more aggressive treatment earlier, such as doublet or triplet therapy, which includes androgen deprivation therapy as the backbone plus new-generation hormonal agents with or without concurrent chemotherapy, to prolong survival [3-5]. Therefore, actively and accurately identifying these patients becomes paramount.

Multi-modality evaluations are required to identify high-risk or high-volume patients. These evaluations should at least include serum prostate-specific antigen level, pathology features with Gleason score, whole-body bone scan (WBBS), computed tomography (CT), and/or magnetic resonance imaging (MRI). To achieve this, clinicians should undertake information retrieval (IR) and risk assessment (RA) tasks. IR includes determining the TNM stage, detecting bone metastasis, counting axial and non-axial bone metastatic sites, and identifying visceral metastasis and site count. This information reflects tumor burden and extensiveness [4]. Beyond IR, prognostic RA requires integrating this information to accomplish complex clinical reasoning based on well-established criteria such as LATITUDE and CHAARTED [3,4].

In daily practice, clinicians should extract information from various sources such as the hospital information system, picture archiving and communication system, laboratory information system, and pathology information management system. However, toggling between these information systems is time-consuming and prone to errors. Therefore, developing a clinical decision-support system (CDSS) that integrates IR and RA capabilities as a reminder could facilitate clinical workflow and reduce errors.

Traditional approaches include keyword searches or building specific natural language processing (NLP) models to accomplish the task. Research has shown that deep learning for NLP approaches, can achieve notable IR results from unstructured cancer pathology reports [6,7]. However, these specific NLP models require large amounts of high-quality training data and human effort for annotation, making the process costly and time-consuming. Recently, a vast array of pre-trained general-purpose large language models (LLMs) has been proposed and has advanced significantly [8]. These models are pre-trained on immense amounts of data, which enables them to perform a wide array of tasks without specific training. This versatility makes them highly cost-effective, as a single general-purpose model might be comparable to multiple specialized ones.

Although general-purpose LLMs were not specifically designed for medical applications, some studies have demonstrated their acceptable performance in IR and RA tasks, however other investigations indicate that LLMs may still be unreliable for broader clinical RAs [9-15]. Such studies typically focus on the model's performance in only one specific task. To our knowledge, there has been scant research that conducts comparative analyses assessing their capabilities in multifaceted IR and RA tasks across multiple LLMs within the medical field.

To address this research gap, this study aims to conduct an in-depth comparison of four cutting-edge LLMs: ChatGPT-4-turbo, Claude-3-opus, Gemini-pro-1.0, and ChatGPT-3.5-turbo. We will evaluate their performance across various IR and RA tasks derived from multi-modality PC work-up reports. This comparison

The World Journal of MEN's HEALTH

seeks to elucidate differences in performance across multifaceted outcomes, providing a comprehensive assessment of each model's capabilities.

MATERIALS AND METHODS

The study flow diagram is illustrated in Fig. 1. The complete methods were detailed in the Supplement Methods 1.

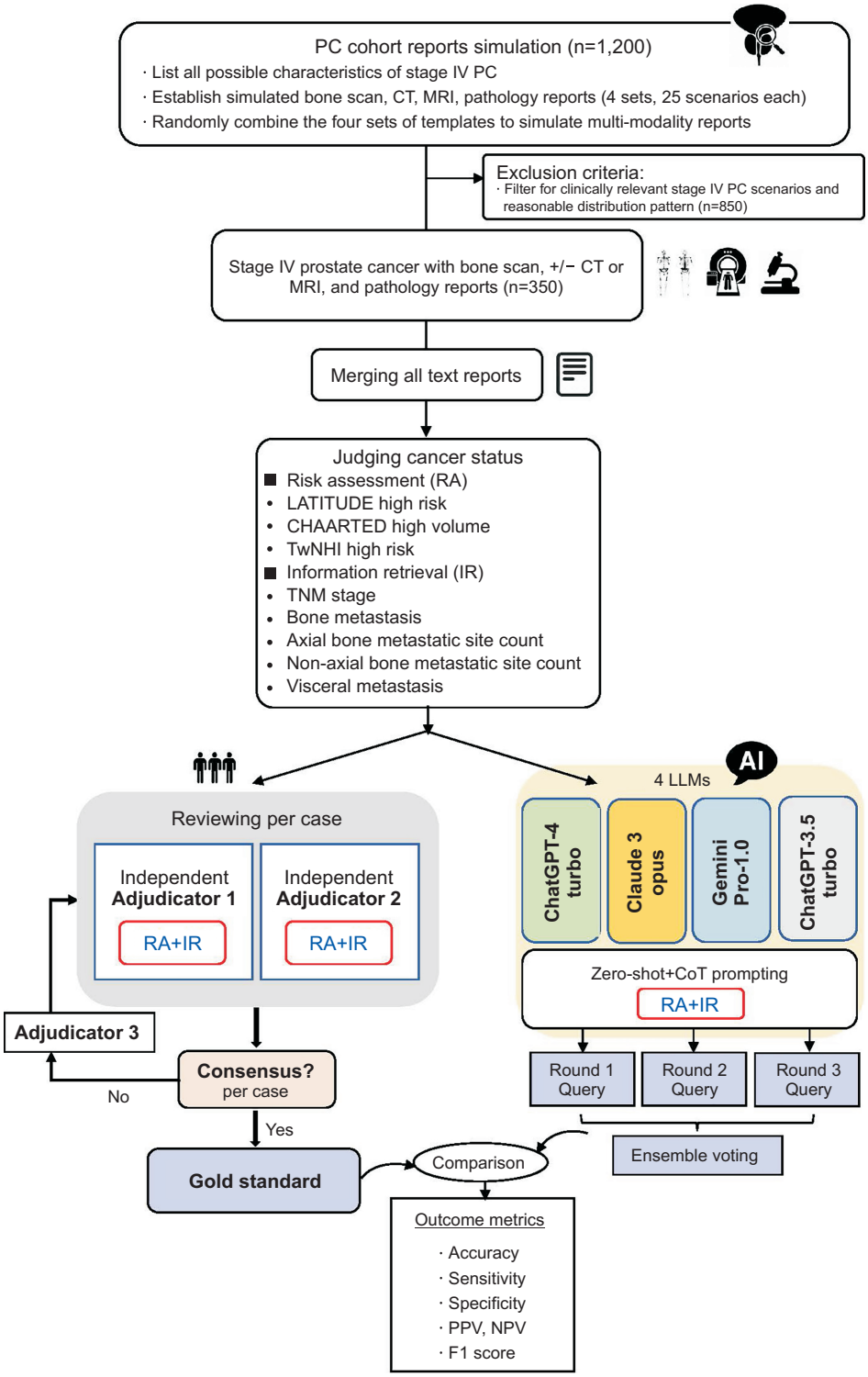

Fig. 1. Study flow diagram. PC: prostate cancer, CT: computed tomography, MRI: magnetic resonance imaging, RA: risk assessment, IR: information retrieval, TNM: tumor, node, metastasis, LLM: large language model, LATITUDE: a randomized, double-blind, comparative study of abiraterone, CHAARTED: chemohormonal therapy *versus* androgen ablation randomized trial for extensive disease in prostate cancer, TwNHI: Taiwan National Health Insurance, CoT: chain-of-thought, PPV: positive predictive value, NPV: negative predictive value.

1. Patient cohort and multi-modality reports simulation

This study simulates 350 stage IV PC patients' work-up text reports, including Bone Scan, CT, MRI, and pathology reports. We ensured that the simulated patient cohort includes various combinations and distributions that closely resemble real-world scenarios. The detailed simulation process described in the Supplement Methods 1.

2. Outcome measurement

We focused on 10 outcomes: 7 IR tasks and 3 RA tasks, analyzing TNM staging, and detection and quantification of bone (axial and non-axial) and visceral metastases from multi-modality PC work-up reports. The RA tasks evaluated metastatic hormone-sensitive PC (mHSPC) using three prognostic risk criteria: LATITUDE high-risk, CHAARTED high-volume, and TwNHI high-risk defined by Taiwan's Health Insurance system.

3. Adjudicators consensus as the gold standard

To establish gold standards for outcomes, two experienced urologists independently assessed the outcomes using integrated imaging and pathological text reports. Discrepancies were resolved through consensus with a third adjudicator.

4. General-purpose pre-trained large language models

Four general-purpose LLMs were used for evaluation: ChatGPT-4-turbo, Claude-3-opus, Gemini Pro 1.0, and ChatGPT-3.5-turbo referenced for assessments and benchmarks.

5. Prompting strategy

We used zero-shot chain-of-thought (ZS-COT) techniques to enhance reasoning and accuracy. Each case began with integrated multi-modality work-up text reports, followed by structured prompts including "Let's think step-by-step" to trigger detailed reasoning.

6. Study design

1) Single-round, repeated 3-round queries, and ensemble voting

Procedures for single-round and repeated 3-round queries: We evaluated each LLM by querying 350 cases using the API. To address potential inconsistencies in LLM outputs, the single-round query procedure was repeated three times for each LLM, measuring consistency with the intraclass correlation coefficient (ICC). Macro-average accuracy was calculated as the arithmetic mean of the accuracy rates from these rounds [16]. The ensemble voting (EV) result was determined by aggregating the majority answer from the three-round results for each case [17].

7. Subgroup analysis

We conducted subgroup analyses evaluating T stages from T2 to T4, M stages including M1a, M1b, and M1c, as well as axial and non-axial bone metastatic site counts using a 3-class classification system: 0, 1–3, and ≥4.

8. Statistical analysis and performance metrics

We used six metrics: accuracy, sensitivity, specificity, positive predictive value (PPV), negative predict value (NPV), and F1-score. Descriptive statistics were reported with interquartile range (IQR) and proportions. Inter-rater reliability was assessed with Cohen's kappa coefficient, and consistency of LLM outputs was evaluated using ICC. Performance comparisons among LLMs were conducted by calculating odds ratios (ORs) from univariate logistic regression, with the least performing model as the reference. All statistical tests were two-tailed, with significance set at $p<0.05$ within a 95% confidence interval (CI).

9. Ethics statement

No human participants were involved in the study. According to the Regulations on Human Trials by the Ministry of Health and Welfare in Taiwan, IRB Board only handles matters related to human subjects and does not accept applications for ethical review when the study does not involve human participants. Also, in accordance with the IRB Review Regulations from Ethics Center of National Taiwan University Hospital (https://www.ntuh.gov.tw/RECO/Fpage.action?muid=5018&fid=5536), the study outside the scope of human research does not need to be submitted to the Research Ethics Committee as per the above regulations. All experiments were performed in accordance with relevant guidelines and regulations.

The World Journal of MEN's HEALTH

RESULTS

The complete results were detailed in the Supplement Results 2.

1. Simulated population characteristics

Among 350 simulated stage IV PC cases, 222 (63.4%) were in T3, and 180 (51.4%) in N1 stage (Table 1). RA classified 115 (32.8%) as LATITUDE high risk, 128 (36.6%) as CHAARTED high volume, and 94 (26.9%) as TwNHI high risk. Adjudicators' inter-rater reliability (kappa) ranged from 0.78 to 0.90 (Supplement Table 3).

2. Accuracy and consistency comparison across 4 LLMs

The accuracy, consistency (ICC), macro-average, and EV accuracy of three single-round queries for 10 tasks across 4 LLMs are detailed in Table 2. ChatGPT-4-turbo showed the best performance in the LATITUDE RA task, with the highest single-round accuracy (89.6%–91.3%), macro-average accuracy (90.1%), EV accuracy (90.1%), and ICC (0.86). Claude-3-opus and Gemini Pro 1.0 followed, while ChatGPT-3.5-turbo had the lowest scores. The rankings were consistent in CHARTED and TwNHI RA.

In TNM staging IR tasks, the performances were similar between 4 LLMS with ICCs ranged from 0.8 to 1.0 and accuracies from 85.7% to 94.1%. For metastases, ICCs were high (>0.8), but Gemini and ChatGPT-3.5 were less consistent in detecting visceral metastases (Table 2).

3. Accuracy comparison between single-round, macro-average and ensemble voting

Table 3 and Fig. 2 show the probability of EV be-

Table 1. Simulation of the stage IV prostate cancer patients cohort and adjudicators' consensus on disease status

	Number (%)
Simulation	
Stage IV prostate cancer cohort	
Total count	350 (100)
Simulated staging multi-modality reports	
Tc99m whole-body bone scan	350 (100)
Computed tomography	194 (55.4)
Magnetic resonance imaging	225 (64.3)
Pathology reports	350 (100)
Adjudicators' consensus in information retrieval[a]	
TNM staging (AJCC 8th ed.)	
T stage	
T2	48 (13.7)
T3	222 (63.4)
T4	80 (22.9)
N stage	
N0	170 (48.6)
N1	180 (51.4)
M stage	
M0	136 (38.9)
M1a	12 (3.4)
M1b	184 (52.6)
M1c	18 (5.1)
Bone metastasis	
Present	198 (56.6)
Abscent	152 (43.4)
Axial bone metastatic site count	
0	165 (47.1)
1–3	63 (18.0)
≥4	122 (34.9)
Non-axial bone metastatic site count	
0	201 (57.4)
1–3	66 (18.0)
≥4	83 (23.7)
Visceral metastasis	
Present	18 (5.1)
Abscent	332 (94.9)
Visceral metastatic site	
Lung	13 (3.7)
Liver	5 (1.4)
Others	3 (0.9)
Adjudicators' consensus in risk assessment[a]	
LATITUDE high risk	
Yes	115 (32.9)
No	235 (67.1)
CHAARTED high volume	
Yes	128 (36.6)
No	222 (63.4)
TwNHI high risk	
Yes	94 (26.9)
No	256 (73.1)

Values are presented as number (%).
AJCC: American Joint Committee on Cancer, LATITUDE: a randomized, double-blind, comparative study of Abiraterone, CHAARTED: chemohormonal therapy *versus* androgen ablation randomized trial for extensive disease in prostate cancer, TwNHI: Taiwan National Health Insurance.
[a]Consensus as the gold standard.

Table 2. Accuracy comparison between single-round queries, macro-average, and ensemble voting with output consistency across four LLMs in 10 risk assessment and information retrieval tasks

	Prediction target	Repeated single-round queries					Ensemble voting
		Round 1	Round 2	Round 3	Consistency (ICC)	Macro-average	
		Accuracy (%)	Accuracy (%)	Accuracy (%)		Accuracy (%)	Accuracy (%)
Risk assessment							
LATITTUDE high risk	Binary						
ChatGPT 4 turbo		89.6	89.6	91.3	0.86	90.1	90.1
Claude 3 opus		74.2	76.5	76.8	0.69	75.8	76.8
Gemini Pro 1.0		61.7	63.2	60.9	0.44	61.9	61.2
ChatGPT 3.5 turbo		36.2	34.5	35.9	0.44	35.6	35.1
CHAARTED high volume	Binary						
ChatGPT 4 turbo		91.3	90.7	90.7	0.93	90.9	90.7
Claude 3 opus		87.0	88.1	88.4	0.85	87.8	88.1
Gemini Pro 1.0		77.1	75.1	77.4	0.63	76.5	77.7
ChatGPT 3.5 turbo		49.0	49.9	50.7	0.58	49.9	49.3
TwNHI high risk	Binary						
ChatGPT 4 turbo		92.2	90.4	87.8	0.76	90.1	91.6
Claude 3 opus		84.3	80.9	82.0	0.66	82.4	84.1
Gemini Pro 1.0		60.3	60.0	62.3	0.48	60.9	62.9
ChatGPT 3.5 turbo		33.3	31.9	33.9	0.41	33.0	31.9
Information retrieval							
T stage	3-class						
ChatGPT 4 turbo		90.1	90.1	89.7	0.91	90.0	90.1
Claude 3 opus		91.0	89.6	93.0	0.92	91.2	93.0
Gemini Pro 1.0		93.2	92.6	91.0	0.97	92.3	92.6
ChatGPT 3.5 turbo		91.2	93.6	93.6	0.89	92.8	91.7
N stage	Binary						
ChatGPT 4 turbo		94.3	94.5	93.5	1.00	94.1	94.2
Claude 3 opus		93.9	93.1	93.9	0.99	93.6	93.9
Gemini Pro 1.0		94.1	93.0	92.4	0.95	93.2	94.0
ChatGPT 3.5 turbo		90.8	92.3	92.3	0.92	91.8	93.3
M stage	4-class						
ChatGPT 4 turbo		90.1	90.1	89.7	0.84	90.0	90.0
Claude 3 opus		90.6	90.2	91.0	0.96	90.6	90.9
Gemini Pro 1.0		85.7	86.2	85.2	0.83	85.7	90.8
ChatGPT 3.5 turbo		84.1	87.1	86.5	0.80	85.9	87.4
Bone metastasis	Binary						
ChatGPT 4 turbo		89.9	89.8	92.1	0.89	90.6	91.3
Claude 3 opus		93.9	92.5	93.0	0.93	93.1	93.3
Gemini Pro 1.0		91.9	92.5	90.4	0.88	91.6	92.5
ChatGPT 3.5 turbo		82.0	82.0	81.7	0.93	81.9	82.0
Axial-bone metastatic site count	3-class						
ChatGPT 4 turbo		84.1	85.0	86.0	0.56	85.0	87.1
Claude 3 opus		85.8	88.1	89.6	0.94	87.8	88.4
Gemini Pro 1.0		85.5	86.7	85.8	0.90	86.0	88.7
ChatGPT 3.5 turbo		71.5	74.4	71.7	0.87	72.6	74.0
Non-axial bone metastatic site count	3-class						
ChatGPT 4 turbo		74.6	73.4	77.2	0.82	75.1	75.6
Claude 3 opus		78.8	77.7	75.1	0.88	77.2	78.8
Gemini Pro 1.0		80.0	79.7	75.7	0.80	78.5	81.4

Table 2. Continued

	Prediction target	Repeated single-round queries					Ensemble voting
		Round 1	Round 2	Round 3	Consistency (ICC)	Macro-average	
		Accuracy (%)	Accuracy (%)	Accuracy (%)		Accuracy (%)	Accuracy (%)
ChatGPT 3.5 turbo		63.2	61.4	62.9	0.82	62.5	64.0
Visceral metastasis	Binary						
ChatGPT 4 turbo		97.4	96.8	95.6	0.86	96.6	96.8
Claude 3 opus		92.2	95.1	94.8	0.83	94.0	94.8
Gemini Pro 1.0		78.8	80.9	83.2	0.59	83.2	84.1
ChatGPT 3.5 turbo		76.2	75.9	77.7	0.72	76.6	77.1

LLM: large language models, ICC: intraclass correlation coefficient, LATITUDE: a randomized, double-blind, comparative study of Abiraterone, CHAARTED: chemohormonal therapy *versus* androgen ablation randomized trial for extensive disease in prostate cancer, TwNHI: Taiwan National Health Insurance.

Table 3. Probability of ensemble voting achieving non-inferior accuracy compared to single-round queries

Category	LLMs	EV non-inferior count	Total comparison	EV non-inferior probability (%)
Risk assessment	ChatGPT 4 turbo	6	9	66.7
	Claude 3 opus	7	9	77.8
	Gemini Pro 1.0	7	9	77.8
	ChatGPT 3.5 turbo	4	9	44.4
Information retrieval	ChatGPT 4 turbo	14	21	66.7
	Claude 3 opus	17	21	81.0
	Gemini Pro 1.0	19	21	90.5
	ChatGPT 3.5 turbo	17	21	81.0

LLM: large language model, EV: ensemble voting.
Non-inferior comparison: The accuracy of EV greater than or equal to that of a single-round query.

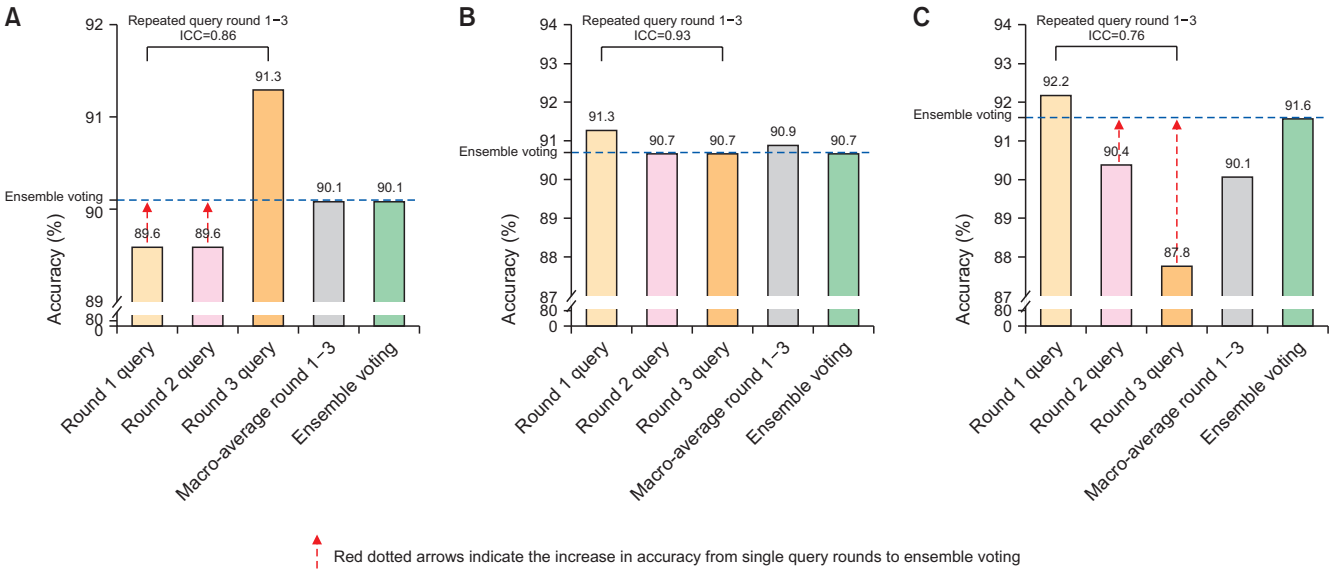

Fig. 2. Accuracy comparison of ChatGPT-4 between 3 repeated single-round queries, macro-average, and ensemble voting. This figure illustrates the accuracy performance of single query rounds in comparison to the ensemble voting method across three different risk assessments: (A) LATITUDE high risk, (B) CHAARTED high volume, (C) TwNHI high risk. The figure demonstrates that ensemble voting consistently presents a higher likelihood of achieving non-inferior accuracy compared to single-round queries. LATITUDE: a randomized, double-blind, comparative study of abiraterone, CHAARTED: chemohormonal therapy *versus* androgen ablation randomized trial for extensive disease in prostate cancer, TwNHI: Taiwan National Health Insurance, ICC: intraclass correlation coefficient.

ing non-inferior to single-round queries greater than 50% across various LLMs in both RA and IR tasks. It showed that compared to single-round queries, the EV accuracy demonstrated non-inferior probabilities ranged from 44.4% to 77.8 % in the RA task and 66.7% to 90.5% in the IR tasks for the 4 LLMs, outperforming macro-average accuracy in most outcomes.

4. Performance evaluation of 4 LLMs based on EV results

Fig. 3 shows an OR-based accuracy comparison from EV results. Fig. 4 and Table 4 present six performance metrics for four LLMs across ten RA and IR tasks. Minimal differences were found in IR performance, with ChatGPT-3.5-turbo lagging. Significant disparities in RA performance ranked the models: ChatGPT-4-turbo, Claude-3-opus, Gemini-Pro-1.0, ChatGPT-3.5-turbo, respectively.

Regarding RA, ChatGPT-4-turbo showed superior performance in LATITUDE, CHAARTED, TwNHI high-risk criteria with the highest EV accuracy, specificity, PPV, and F1 score. In the LATITUDE high-risk task, ChatGPT-4 ranked first, followed by Claude-3-opus and Gemini Pro 1.0, with ORs of 16.9 (11.2–25.7), 5.3

Ensemble voting	Accuracy (%)	OR	95% CI	p value
Risk assessment				
LATITTUDE high risk				
ChatGPT 4 turbo	90.1	16.93	11.16–25.71	<0.001
Claude 3 opus	76.8	5.32	3.83–7.39	<0.001
Gemini Pro 1.0	61.2	2.92	2.13–3.97	<0.001
ChatGPT 3.5 turbo	35.1	ref		
CHAARTED high volume				
ChatGPT 4 turbo	90.7	10.07	6.61–15.33	<0.001
Claude 3 opus	88.1	7.63	5.18–11.26	<0.001
Gemini Pro 1.0	77.7	3.58	2.58–4.98	<0.001
ChatGPT 3.5 turbo	49.3	ref		
TwNHI high risk				
ChatGPT 4 turbo	91.6	23.28	14.95–36.24	<0.001
Claude 3 opus	84.1	11.26	7.81–16.25	<0.001
Gemini Pro 1.0	62.9	3.62	2.64–4.96	<0.001
ChatGPT 3.5 turbo	31.9	ref		
Information retrival				
T stage				
ChatGPT 4 turbo	90.1	0.83	0.42–1.63	0.584
Claude 3 opus	93.0	1.21	0.58–2.53	0.605
Gemini Pro 1.0	92.6	1.14	0.55–2.38	0.722
ChatGPT 3.5 turbo	91.7	ref		
N stage				
ChatGPT 4 turbo	94.5	1.17	0.59–2.32	0.655
Claude 3 opus	93.9	1.10	0.55–2.21	0.789
Gemini Pro 1.0	94.0	1.11	0.54–2.30	0.779
ChatGPT 3.5 turbo	93.3	ref		
M stage				
ChatGPT 4 turbo	90.0	1.30	0.77–2.19	0.330
Claude 3 opus	90.9	1.45	0.83–2.52	0.193
Gemini Pro 1.0	90.8	1.43	0.75–2.70	0.275
ChatGPT 3.5 turbo	87.4	ref		
Bone metastasis				
ChatGPT 4 turbo	91.0	2.30	1.45–3.66	<0.001
Claude 3 opus	93.3	3.07	1.85–5.08	<0.001
Gemini Pro 1.0	92.5	2.69	1.66–4.37	<0.001
ChatGPT 3.5 turbo	82.0	ref		
Axial_bone metastasis				
ChatGPT 4 turbo	87.1	2.45	1.61–3.74	<0.001
Claude 3 opus	88.4	2.75	1.83–4.13	<0.001
Gemini Pro 1.0	88.7	2.91	1.92–4.40	<0.001
ChatGPT 3.5 turbo	74.0	ref		
Non-axial bone metastasis				
ChatGPT 4 turbo	75.4	1.74	1.25–2.43	<0.001
Claude 3 opus	78.8	2.09	1.49–2.94	<0.001
Gemini Pro 1.0	81.4	2.61	1.83–3.73	<0.001
ChatGPT 3.5 turbo	64.0	ref		
Visceral metastasis				
ChatGPT 4 turbo	96.8	9.02	4.70–17.29	<0.001
Claude 3 opus	94.8	5.40	3.15–9.23	<0.001
Gemini Pro 1.0	84.1	1.57	1.07–2.23	0.021
ChatGPT 3.5 turbo	77.1	ref		

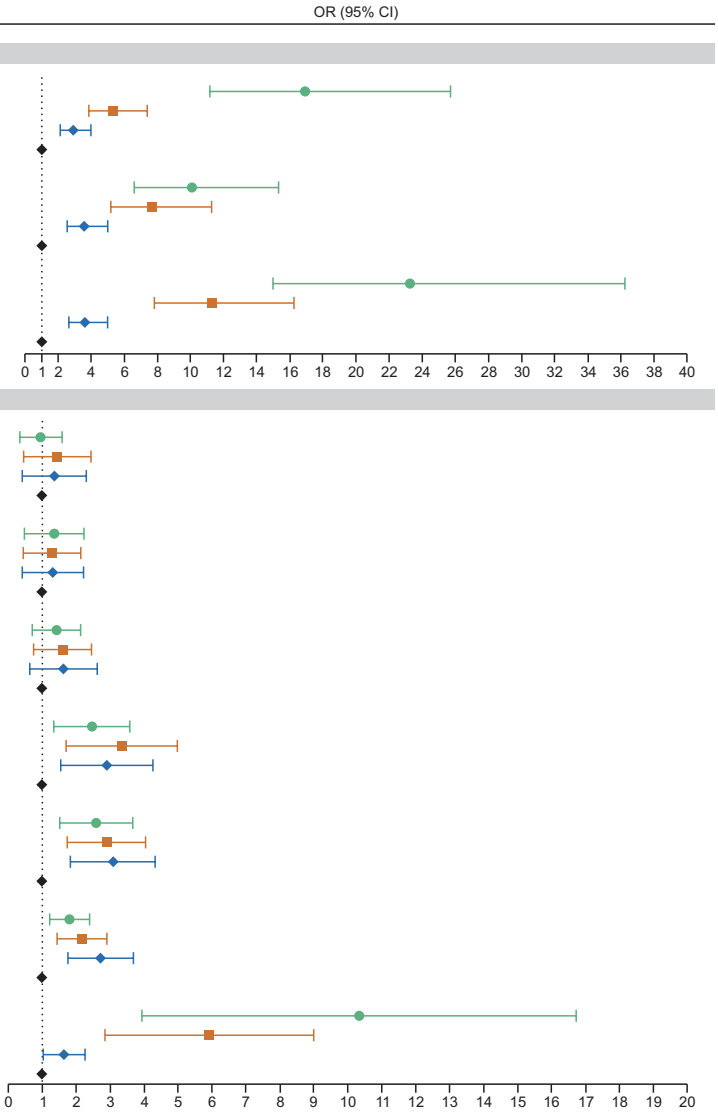

Fig. 3. Comparison of ensemble voting accuracy between large language models (LLMs) with forest plot. The figure compares the accuracy of four LLMs: ChatGPT-4-turbo, Claude-3-opus, Gemini Pro 1.0, and ChatGPT-3.5-turbo (as the reference model) based on ensemble voting result in 3 risk assessment and 7 information retrieval tasks. OR: odds ratio, CI: confidence interval, LATITUDE: a randomized, double-blind, comparative study of Abiraterone, CHAARTED: chemohormonal therapy *versus* androgen ablation randomized trial for extensive disease in prostate cancer, TwNHI: Taiwan National Health Insurance.

The World Journal of MEN's HEALTH

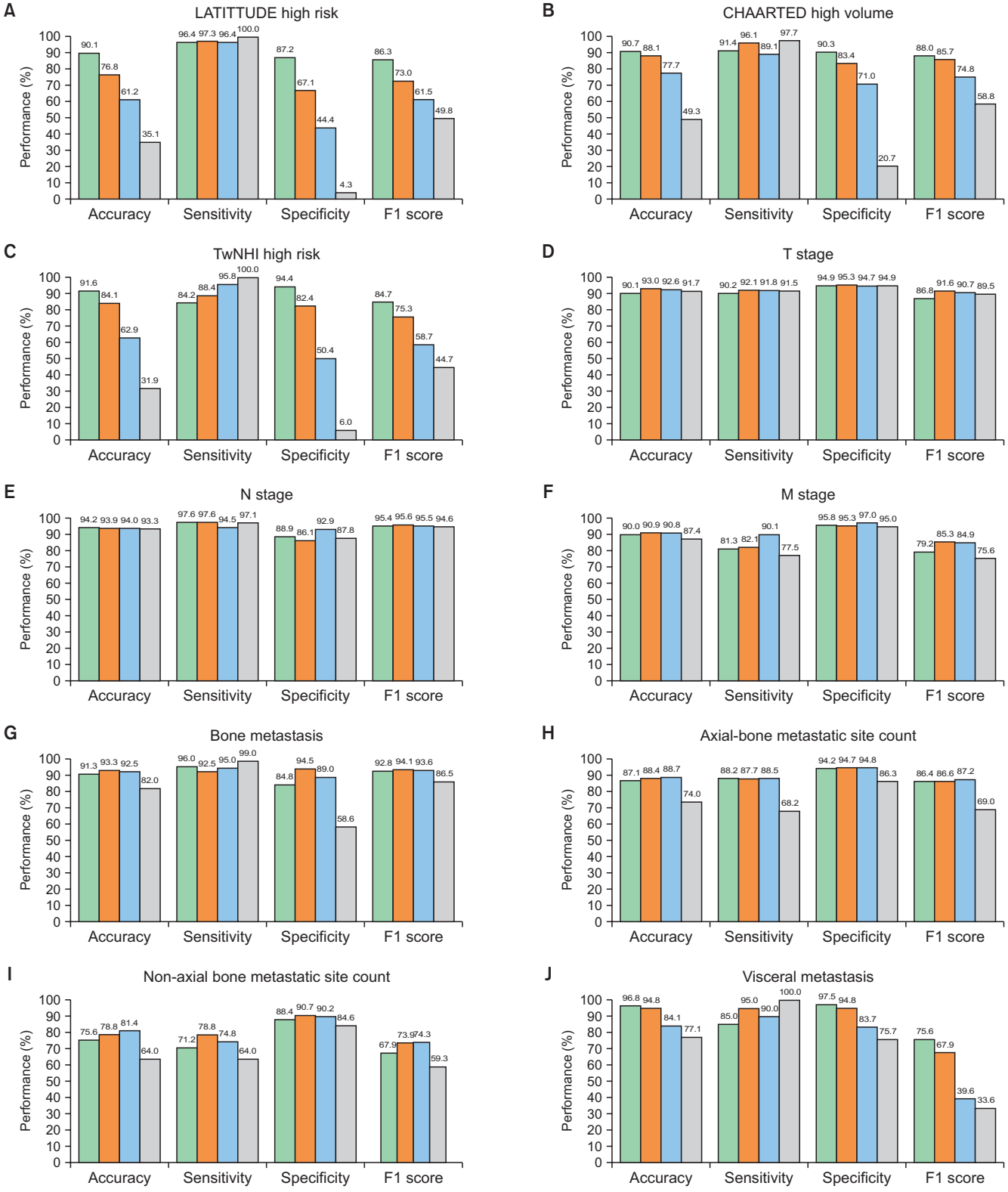

Fig. 4. Comparative performance metrics of four large language models (LLMs) across ten risk assessment and information retrieval tasks. Comparative performance metrics—accuracy, sensitivity, specificity, and F1 score—for four LLMs: ChatGPT-4-turbo, Claude-3-opus, Gemini Pro 1.0, and ChatGPT-3.5-turbo. The analysis spans risk assessment tasks for LATITUDE high-risk (A), CHAARTED high-volume (B), and TwNHI high-risk (C); TNM staging tasks for T stage (D), N stage (E), and M stage (F); and metrics for bone metastasis overall (G), axial-bone metastatic site count (H), non-axial bone metastatic site count (I), and visceral metastasis (J). LATITUDE: a randomized, double-blind, comparative study of Abiraterone, CHAARTED: chemohormonal therapy *versus* androgen ablation randomized trial for extensive disease in prostate cancer, TwNHI: Taiwan National Health Insurance.

Table 4. Performance evaluation of 4 LLMs in 10 risk assessment and information retrieval tasks: ensemble voting results

Category	LLM	Ensemble voting					
		Accuracy (%)	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	F1-score (%)
Risk assessment							
LATITTUDE high risk							
	ChatGPT 4 turbo	90.1	96.4	87.2	78.1	98.1	86.3
	Claude 3 opus	76.8	97.3	67.1	58.4	98.1	73.0
	Gemini Pro 1.0	61.2	96.4	44.4	45.1	96.3	61.5
	ChatGPT 3.5 turbo	35.1	100.0	4.3	33.1	100.0	49.8
CHAARTED high volume							
	ChatGPT 4 turbo	90.7	91.4	90.3	84.8	94.7	88.0
	Claude 3 opus	88.1	96.1	83.4	77.4	97.3	85.7
	Gemini Pro 1.0	77.7	89.1	71.0	64.4	91.7	74.8
	ChatGPT 3.5 turbo	49.3	97.7	20.7	42.1	93.8	58.8
TwNHI high risk							
	ChatGPT 4 turbo	91.6	84.2	94.4	85.1	94.0	84.7
	Claude 3 opus	84.1	88.4	82.4	65.6	94.9	75.3
	Gemini Pro 1.0	62.9	95.8	50.4	42.3	96.9	58.7
	ChatGPT 3.5 turbo	31.9	100.0	6.0	28.8	100.0	44.7
Information retrieval							
T stage							
	ChatGPT 4 turbo	90.1	90.2	94.9	84.6	93.4	86.8
	Claude 3 opus	93.0	92.1	95.3	91.3	95.5	91.6
	Gemini Pro 1.0	92.6	91.8	94.7	90.8	95.4	90.7
	ChatGPT 3.5 turbo	91.7	91.5	94.9	88.5	94.8	89.5
N stage							
	ChatGPT 4 turbo	94.2	97.6	88.9	93.3	96.0	95.4
	Claude 3 opus	93.9	97.6	86.1	93.7	94.4	95.6
	Gemini Pro 1.0	94.0	94.5	92.9	96.5	89.0	95.5
	ChatGPT 3.5 turbo	93.3	97.1	87.8	92.2	95.3	94.6
M stage							
	ChatGPT 4 turbo	90.0	81.3	95.8	77.9	97.1	79.2
	Claude 3 opus	90.9	82.1	95.3	89.3	96.2	85.3
	Gemini Pro 1.0	90.8	90.1	97.0	81.4	95.5	84.9
	ChatGPT 3.5 turbo	87.4	77.5	95.0	74.6	94.9	75.6
Bone metastasis							
	ChatGPT 4 turbo	91.3	96.0	84.8	89.7	93.9	92.8
	Claude 3 opus	93.3	92.5	94.5	95.9	90.1	94.1
	Gemini Pro 1.0	92.5	95.0	89.0	92.2	92.8	93.6
	ChatGPT 3.5 turbo	82.0	99.0	58.6	76.7	97.7	86.5
Axial-bone metastatic site count							
	ChatGPT 4 turbo	87.1	88.2	94.2	86.6	92.9	86.4
	Claude 3 opus	88.4	87.7	94.7	86.0	94.1	86.6
	Gemini Pro 1.0	88.7	88.5	94.8	87.1	94.2	87.2
	ChatGPT 3.5 turbo	74.0	68.2	86.3	70.7	87.1	69.0
Non-axial bone metastatic site count							
	ChatGPT 4 turbo	75.6	71.2	88.4	73.8	87.3	67.9
	Claude 3 opus	78.8	78.8	90.7	77.0	88.6	73.9
	Gemini Pro 1.0	81.4	74.8	90.2	75.3	90.0	74.3
	ChatGPT 3.5 turbo	64.0	64.0	84.6	61.4	81.8	59.3

Table 4. Continued

Category	LLM	Ensemble voting					
		Accuracy (%)	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	F1-score (%)
Visceral metastasis							
	ChatGPT 4 turbo	96.8	85.0	97.5	68.0	99.1	75.6
	Claude 3 opus	94.8	95.0	94.8	52.8	99.7	67.9
	Gemini Pro 1.0	84.1	90.0	83.7	25.4	99.3	39.6
	ChatGPT 3.5 turbo	77.1	100.0	75.7	20.2	100.0	33.6

LLM: large language model, PPV: positive predictive value, NPV: negative predictive value, LATITUDE: a randomized, double-blind, comparative study of Abiraterone, CHAARTED: chemohormonal therapy *versus* androgen ablation randomized trial for extensive disease in prostate cancer, TwNHI: Taiwan National Health Insurance.

(5.2–11.3), and 2.9 (2.1–3.97), respectively, compared to ChatGPT-3.5. These rankings were consistent across all RA tasks, with ChatGPT-4 significantly outperforming Claude-3. It is noteworthy that the sensitivity and NPV were exceptionally high across all RA tasks, ranging from 85% to 100% and 91% to 100%, respectively.

For TNM staging, no significant differences were found in accuracy, sensitivity, specificity, NPV, PPV, and F1 score among the four models. The ORs for EV accuracy in T-stage classification were similar across models.

Regarding bone metastasis detection, ChatGPT-4, Claude-3-opus, and Gemini Pro 1.0 outperformed ChatGPT-3.5. ORs for accuracy were 2.45 (1.61–3.74), 2.75 (1.83–4.13), and 2.69 (1.66–4.37) for ChatGPT-4, Claude-3-opus and Gemini Pro 1.0, respectively. ChatGPT-3.5 had notably lower specificity and PPV. In detecting visceral metastasis, ChatGPT-3.5 and Gemini-Pro-1.0 showed poorer specificity and PPV compared to ChatGPT-4 and Claude-3-opus.

5. Subgroup analysis

We further compared ChatGPT-4-turbo, Claude-3-opus, and Gemini Pro 1.0 performance across each class in tumor, metastasis, and bone metastatic site counts. The majority of parameters showed good performance across subgroups in these three models. However, sensitivity and PPV are low in M1a and non-axial bone metastatic site counts of 1–3 and ≥4 (Table 5, Supplement Table 4, 5).

DISCUSSION

This study focuses on IR and RA tasks for stage IV PC. We selected four general-purpose LLMs for comparison. In terms of IR performance, the four models showed minimal differences, with ChatGPT-3.5-turbo slightly lagging behind. However, there were significant differences in RA performance, with the ranking as follows: ChatGPT-4-turbo, Claude-3-opus, Gemini-Pro-1.0, ChatGPT-3.5-turbo, respectively. ChatGPT-4-turbo outperformed the other three LLMs in all three RA tasks across three metrics: accuracy, F1-score, and consistency of repeated queries (ICC). Consequently, we conclude that ChatGPT-4-turbo has the best performance for our tasks. Moreover, our data demonstrated that the general-purpose ChatGPT-4 model, without any domain-specific training, can still achieve promising and satisfactory performance in RA for advanced PC. More specifically, across the three different RA tasks, ChatGPT-4 demonstrated satisfactory performance in terms of sensitivity and NPV. This suggests that ChatGPT-4 could be effectively utilized as a screening tool within CDSS to rule out non-high-risk PC patients. To our knowledge, this is the first study to simultaneously compare four different LLMs across ten IR and RA tasks in the medical field. Our findings indicate that general-purpose LLMs could be considered a viable option for hospitals to adopt within CDSS.

This study conducted comparisons across seven different IR tasks. While most tasks showed satisfactory performance, the 3-class classification tasks for axial and non-axial bone metastatic site counts performed only fairly. Our dataset consists of pathology and image reports that are unstructured and presented in free-text format. Tasks such as determining TNM stages, identifying the presence or absence of bone and visceral metastases (binary outcome), and counting axial and non-axial bone metastatic sites (multiclass classification) require different levels of AI cognitive abilities. The retrieval of TNM stages involves search-

Table 5. Subgroup analysis of ChatGPT-4's performance in information retrieval tasks based on ensemble voting results

	Multiclass classification	Ensemble voting					
		Accuracy (%)	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	F1 score (%)
Information retrieval							
T stage[a]	3-class	93.4	90.2	94.9	84.6	93.4	86.8
T2		95.0	95.8	94.9	71.9	99.4	82.1
T3		92.6	91.5	94.4	96.7	86.1	94.1
T4		92.6	83.3	95.5	85.1	94.8	84.2
M stage[a]	4-class	95.0	81.3	95.8	77.9	97.1	79.2
M0		91.3	79.3	96.1	89.0	92.1	83.9
M1a		95.8	60.0	97.1	42.9	98.5	50.0
M1b		94.5	96.6	91.0	94.5	98.5	95.6
M1c		98.3	89.5	98.9	85.0	99.3	87.2
Axial-bone metastatic site count[a]	3-class	91.4	88.2	94.2	86.6	92.9	86.4
0		91.0	84.4	98.5	98.4	85.0	90.8
1–3		88.2	95.5	85.8	68.1	98.4	79.5
≥4		95.0	84.6	98.1	93.2	95.5	88.7
Non-axial bone metastatic site count[a]	3-class	83.7	71.2	88.4	73.8	87.3	67.9
0		87.3	86.4	88.7	92.0	81.4	89.1
1–3		78.3	79.6	78.1	38.6	95.7	52.0
≥4		85.5	47.6	98.4	90.9	84.7	62.5

PPV: positive predictive value, NPV: negative predictive value.
[a]The six metrics of the task category were calculated by macro-average across subgroups.

ing for relevant keywords within the reports. For the counting of bone or visceral metastatic sites, the LLM must interpret terms such as 'multiple,' 'diffuse,' 'probable,' 'unlikely,' 'maybe,' and 'significant.' It is difficult to accurately count and categorize the number of metastatic sites. The vague and uncertain language commonly found in these descriptions can pose challenges for both LLMs and human evaluators.

RA requires high-level cognitive functions, including domain-specific medical knowledge, text report comprehension, IR, and multiple-criteria logical reasoning. Integrating these complex tasks presents a significant challenge for LLMs. In our comparison of three RA tasks across four LLMs, ChatGPT-4 surprisingly outperformed the other models, demonstrating promising and satisfactory results across multiple metrics. Additionally, we intentionally included TwNHI high-risk criteria as a reference outcome. The CHARTED and LATITUDE criteria, widely used in prognostic RAs both in literature and clinical settings, are likely included in the pre-trained datasets for LLMs. However, the TwNHI high-risk criteria, which are specific to Taiwan and documented in traditional Mandarin characters, are unlikely to be incorporated into LLMs' training datasets. Despite this, ChatGPT-4 was able to perform comparably to the LATITUDE and CHART-ED criteria in zero-shot prompting with only brief explanations. This suggests that ChatGPT-4's inherent logical reasoning ability for RA tasks is robust, allowing it to excel even when the criteria are not included in its training data. Our research shows that ChatGPT-4 achieved RA task accuracies of 90.1%, 90.7%, and 91.6%, closely approaching the level of human physicians. Even expert human judgments can vary, as evidenced by inter-rater reliability kappa values of 0.89, 0.84, and 0.89, respectively.

When comparing the performance ranking of the four LLMs, our findings are consistent with the overall model benchmark results for non-medical-specific tasks, as seen in Chatbot Arena [18] and CompassRank [19]. However, according to Anthropic's evaluation report, Claude 3 slightly outperformed ChatGPT-4 in reasoning over test capability (83.1 *vs.* 80.9) [20]. While there may be slight differences in ranking, the overall trend is clear: within just one year, the new generation of ChatGPT-4, Claude-3, and Gemini-1 has significantly surpassed the previous generation ChatGPT-3.5. This indicates that general-purpose LLMs are improving at

The World Journal of MEN's HEALTH

an astonishingly rapid pace.

Similar to our results, Meral et al [13] compared the performance of ChatGPT-4, Gemini, and emergency medicine specialists in ESI triage assessment. They found that ChatGPT-4 performed best in correct triage across all patients, surpassing both Gemini and the emergency medicine specialists. Levkovich et al [21] compared ChatGPT-4 with ChatGPT-3.5 in suicide RA. ChatGPT-4 estimates the likelihood of suicide attempts in a manner akin to evaluations provided by professionals, while ChatGPT-3.5 frequently underestimated suicide risk. A prospective multicenter study evaluated the accuracy of ChatGPT-4 in predicting the ASA score alongside anesthesiologists [12]. They found a high level of concordance between ChatGPT-4 predictions and anesthesiologist evaluations, with a kappa value of 0.858 [13]. In contrast, Raghu et al [14] explored ChatGPT's ability to predict the risk of diabetic retinopathy from clinical data. They discovered that ChatGPT-4 had high consistency and reliability, with an ICC greater than 0.9, but only a fair accuracy range of 0.6 to 0.7. However, Heston and Lewis et al [15] demonstrated inconsistent RA and poor reliability in non-traumatic acute chest pain by ChatGPT-4.

In the subgroup analysis, we found that the sensitivity and PPV of the LLM for M1a (paraaortic lymph node metastases only) and non-axial bone metastasis site count were not as high as for other outcomes, regardless of the model used. This may be attributed to data imbalances within different subgroups, as only 12 patients belonged to M1a, where performance is typically inferior in less common scenarios. Additionally, descriptions of non-axial bone metastases are more complex; accurately counting the numbers and providing correct classifications may be more challenging, even for experienced nuclear radiologists and urologists. Moreover, the multiclass classification may be a more challenging task.

We employed EV methods to overcome the issue of inconsistent output from LLMs. The accuracy of EV has been proven to be superior or equal to that of a single-round query in most circumstances. This approach effectively addresses the instability and variability of the LLM, thereby enhancing its accuracy. Regarding the model's reliability, the consistency of repeated queries (ICC) parallels the accuracy rates of different models. It appears that for LLMs, validity and reliability are aligned; as validity increases, reliability also tends to improve.

Currently, the development of CDSS commonly involves incorporating customized or domain-specific models and algorithms tailored to specific tasks because they commonly take advantages in accuracy and specialization. However, building a domain-specific model or fine-tuning an existing LLM requires substantial amounts of high-quality training data, significant human effort for annotation, and considerable infrastructure investment, including hiring AI professionals. This approach is notably costly and time-consuming. Additionally, fine-tuning LLMs poses potential risks, such as catastrophic forgetting, where the model may lose its ability to perform other tasks [22]. Moreover, it is impractical for a hospital to develop a specific model for each task. In contrast, general-purpose LLMs possess versatility, with a single model capable of handling a wide range of tasks across various domains. Second, they can be easily and rapidly deployed in hospitals without specific training. Third, they evolve quickly, with new versions offering improved performance being released in short time periods. Therefore, if general-purpose language models can achieve satisfactory accuracy, they may be a more practical solution for hospitals than building multiple domain-specific models [23,24].

Our general-purpose LLM approach can be seamlessly integrated into CDSS to facilitate IR and RA tasks. This integration provides healthcare professionals with one-click access, thereby improving healthcare delivery. Once activated, the CDSS operates in the background, compiling multi-modality PC reports into predefined ZS-COT prompts. These prompts are then transmitted through an API to the LLM, which processes and retrieves results from three rounds of queries. The CDSS then calculates the EV results and directly delivers these insights to physicians. The entire process can be fully automated and integrated into existing workflows as a supplementary tool.

However, the potential risks and ethical considerations of implementing AI tools for IR and RA in clinical settings should be approached with caution. Inaccurate IR and RA could mislead clinical decision-making, particularly when physicians overly rely on these tools without incorporating their own judgment. Despite ChatGPT-4-turbo showing optimal performance, with accuracy exceeding 90% in eight out of ten tasks, our analysis identified several types of errors that under-

line the limitations of LLMs. First, errors can occur when there is a discrepancy between different imaging modalities. For instance, LLMs may fail to prioritize CT and MRI findings over WBBSs, despite the higher specificity of the former. Second, in complex cases, such as patients with double cancers or multiple metastases, LLMs may generate conflicting interpretations. Third, regarding bone metastasis, LLMs might overinterpret suspicious lesions in WBBS reports, leading to false positives. This tendency results in high sensitivity (96.0%) but slightly lower specificity (84.8%), culminating in a false positive rate of 15.2%. These findings indicate that the performance of LLMs has not yet reached a fully reliable level, and there is a risk of errors. Consequently, LLMs could only be considered supplementary decision-support tools for physicians, not replacements for clinical judgment. When considering the use of these AI tools, it is essential to clearly disclose their potential error risks to physicians to meet the requirements of responsible AI.

There are several limitations to our study. First, our patient data consisted of simulated text reports, which may not fully represent the true population. Due to privacy and ethical concerns, transmitting the complete real reports of individual patients to any third-party AI model for queries is still prohibited in most institutional review boards. As a result, employing simulated patient data to test and evaluate AI models is a commonly used research method in healthcare and medical research [13,14,25,26]. Nonetheless, we made every effort to simulate the diversity of reports and approximate real-world distribution. Consequently, three adjudicators, who were blind to the study context, did not notice any anomalies while interpreting the reports. Second, our simulated text reports were derived from limited sources, and the descriptions of pathology or imaging reports can vary significantly across different medical facilities and even among individual physicians. As a result, the extrapolation and generalizability of our findings require further evaluation. Third, our research exclusively focused on patients with stage IV PC. Therefore, the promising results observed with ChatGPT-4-turbo in this specific context may not be generalizable to other cancer types or different stages of PC. Further research is necessary to validate these findings in other cancers. Fourth, given the rapid advancements in AI technologies, the findings from this cross-sectional study may only represent the models' performance at this specific point in time and could vary in the future, Fifth, we utilized a zero-shot chain of thoughts as our prompting strategy. We did not explore other prompt strategies, such as one-shot or few-shot approaches.

CONCLUSIONS

This study conclusively shows that ChatGPT-4-turbo is the most effective among the four general-purpose LLMs tested, particularly excelling in RA tasks for stage IV PC with superior accuracy, F1-score, and consistency. Despite the lack of domain-specific training, ChatGPT-4-turbo combined with EV achieved promising results, suggesting its potential utility as a supplementary tool in CDSS for screening high-risk PC patients, thus facilitating clinical workflows. However, the potential risks of inaccurate IR and RA could mislead clinical decision-making. Therefore, implementing LLMs in clinical settings should be approached with caution and accompanied by risk disclosure.

Conflict of Interest

The authors have nothing to disclose.

Funding

None.

Acknowledgements

We would like to acknowledge the efforts of the three independent adjudicators who evaluated thee simulated prostate cancer patients' text reports.

Author Contribution

Conceptualization: SWH, CYT. Data curation: LHY, SWH. Formal analysis: LHY, SWH, CYT. Supervision: CYT, DC. Visualization: CYT. Writing – original draft: LHY, SWH, CYT. Writing – review & editing: CYT, DC.

Supplementary Materials

Supplementary materials can be found via https://doi.org/10.5534/wjmh.240173.

Data Sharing Statement

The datasets generated and/or analyzed during the current study are available from the corresponding author upon reasonable request.

REFERENCES

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2021;71:209-49.
2. Abdollah F, Sun M, Thuret R, Jeldres C, Tian Z, Briganti A, et al. A competing-risks analysis of survival after alternative treatment modalities for prostate cancer patients: 1988-2006. Eur Urol 2011;59:88-95.
3. Fizazi K, Tran N, Fein L, Matsubara N, Rodriguez-Antolin A, Alekseev BY, et al. Abiraterone plus prednisone in metastatic, castration-sensitive prostate cancer. N Engl J Med 2017;377:352-60.
4. Sweeney CJ, Chen YH, Carducci M, Liu G, Jarrard DF, Eisenberger M, et al. Chemohormonal therapy in metastatic hormone-sensitive prostate cancer. N Engl J Med 2015;373:737-46.
5. Kyriakopoulos CE, Chen YH, Carducci MA, Liu G, Jarrard DF, Hahn NM, et al. Chemohormonal therapy in metastatic hormone-sensitive prostate cancer: long-term survival analysis of the randomized phase III E3805 CHAARTED Trial. J Clin Oncol 2018;36:1080-7.
6. Gao S, Young MT, Qiu JX, Yoon HJ, Christian JB, Fearn PA, et al. Hierarchical attention networks for information extraction from cancer pathology reports. J Am Med Inform Assoc 2018;25:321-30.
7. Wu J, Tang K, Zhang H, Wang C, Li C. Structured information extraction of pathology reports with attention-based graph convolutional network. Paper presented at: 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM); 2020 Dec 16-19; Seoul, Korea. p. 2395-402.
8. Naveed H, Khan AU, Qiu S, Saqib M, Anwar S, Usman M, et al. A comprehensive overview of large language models. arXiv 2307.06435 [Preprint]. 2024 [cited 2024 Apr 9]. Available from: https://doi.org/10.48550/arXiv.2307.06435
9. Goel A, Gueta A, Gilon O, Liu C, Erell S, Nguyen LH, et al. LLMs accelerate annotation for medical information extraction. Machine Learning for Health (ML4H); 2023:82-100.
10. Ahsan H, McInerney DJ, Kim J, Potter C, Young G, Amir S, et al. Retrieving evidence from EHRs with LLMs: Possibilities and challenges. arXiv 2309.04550 [Preprint]. 2024 [cited 2024 Mar 3]. Available from: https://doi.org/10.48550/arXiv.2309.04550
11. Change CH, Lucas MM, Lu-Yao G, Yang CC. Classifying cancer stage with open-source clinical large language models. Paper presented at: 2024 IEEE 12th International Conference on Healthcare Informatics (ICHI); 2024 Jun 3-6; Orlando, FL, USA. p. 76-82.
12. Turan E, Baydemir AE, Özcan FG, Şahin AS. Evaluating the accuracy of ChatGPT-4 in predicting ASA scores: a prospective multicentric study ChatGPT-4 in ASA score prediction. J Clin Anesth 2024;96:111475.
13. Meral G, Ateş S, Günay S, Öztürk A, Kuşdoğan M. Comparative analysis of ChatGPT, Gemini and emergency medicine specialist in ESI triage assessment. Am J Emerg Med 2024;81:146-50.
14. Raghu K, S T, C SD, M S, Rajalakshmi R, Raman R. The utility of ChatGPT in diabetic retinopathy risk assessment: a comparative study with clinical diagnosis. Clin Ophthalmol 2023;17:4021-31.
15. Heston TF, Lewis LM. ChatGPT provides inconsistent risk-stratification of patients with atraumatic chest pain. PLoS One 2024;19:e0301854.
16. Liljequist D, Elfving B, Skavberg Roaldsen K. Intraclass correlation - a discussion and demonstration of basic features. PLoS One 2019;14:e0219854.
17. Dietterich TG. Ensemble methods in machine learning. In: Multiple Classifier Systems; Lecture Notes in Computer Science, vol 1857; 2000 Jan 1; Berlin, Heidelberg. Springer; c2000. p. 1-15.
18. Chatbot Arena LLM leaderboard: community-driven evaluation for best LLM and AI chatbots [Internet]. LMSYS; c2024 [cited 2024 Mar 1]. Available from: https://chat.lmsys.org/?leaderboard
19. OpenCompass. CompassRank [Internet]. OpenCompass; c2024 [cited 2024 May 1]. Available from: https://rank.opencompass.org.cn/home
20. Introducing the next generation of Claude [Internet]. Anthropic PBC; c2024 [cited 2024 Mar 4]. Available from: https://www.anthropic.com/news/claude-3-family
21. Levkovich I, Elyoseph Z. Suicide risk assessments through the eyes of ChatGPT-3.5 versus ChatGPT-4: vignette study. JMIR Ment Health 2023;10:e51232.
22. Luo Y, Yang Z, Meng F, Li Y, Zhou J, Zhang Y. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. arXiv 2308.08747 [Preprint]. 2023 [cited 2024 Aug 21]. Available from: https://doi.org/10.48550/arXiv.2308.08747
23. Fedeli A, Beutling N, Laurenzi E, Polini A. Comparison of

general-purpose and domain-specific modelling languages in the IoT domain: A case study from the OMiLAB community. Paper presented at: BIR 2023 Workshops and Doctoral Consortium, 22nd International Conference on Perspectives in Business Informatics Research (BIR 2023); 2023 Sep 13-15; Ascoli Piceno, Italy. p. 145-57.

24. Comparative analysis of custom LLM vs. general-purpose LLM [internet]. CAPER; c2023 [cited 2024 May 1]. Available from: https://gaper.io/custom-llm-vs-general-purpose-llm/
25. Kernberg A, Gold JA, Mohan V. Using ChatGPT-4 to create structured medical notes from audio recordings of physician-patient encounters: comparative study. J Med Internet Res 2024;26:e54419.
26. Levin C, Suliman M, Naimi E, Saban M. Augmenting intensive care unit nursing practice with generative AI: a formative study of diagnostic synergies using simulation-based clinical cases. J Clin Nurs 2024. doi: 10.1111/jocn.17384 [Epub].