Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Patient-Facing AI Chatbots vs Primary Care Physicians: A Standardized Patient Field Experiment in China

Yafei Si
yafei.si@unimelb.edu.au

The University of Melbourne
Shaoqing Gong
Luohe Medical college
Michael Thielscher
UNSW Sydney
Hazel Bateman
UNSW Sydney
Bingqin Li
UNSW Sydney
Shanquan Chen
University of Hong Kong https://orcid.org/0000-0002-4724-4892
Duolao Wang
Liverpool School of Tropical Medicine https://orcid.org/0000-0003-2788-2464
Ruopeng An
Yurun Meng
Xi'an Jiaotong University
Han Zhang
Xi'an Jiaotong University
Suya An
Xi'an Jiaotong University
Jiaqi Zu
RMIT University
Min Su
Inner Mongolia University
Yudong Miao
Zhengzhou University https://orcid.org/0000-0002-8007-1129
Zhongliang Zhou
Xi'an Jiaotong University
Xiaojing Fan
Xi'an Jiaotong University
Limin Mao
UNSW Sydney
Xi Chen
Yale University https://orcid.org/0000-0002-2058-0351
Gang Chen
The University of Melbourne

Article

Keywords: primary health care, patient-facing AI, quality of care, clinical safety, patient-centeredness

Posted Date: June 28th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9988886/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: There is **NO** Competing Interest.

Abstract

Importance

Primary health care (PHC) systems worldwide face persistent workforce shortages, access barriers, and quality gaps. As patients increasingly turn to low-cost easily accessible AI chatbots for health advice, real-world evidence is needed on how patient-facing AI performs compared with PHC physicians.

Objective

To compare patient-facing AI chatbots with PHC physicians in diagnostic accuracy, clinical safety, and patient-centered communication using standardized patients in real-world PHC settings.

Design, Setting, and Participants

This standardized patient field experiment was conducted from February 6 to February 27, 2025, in 62 PHC institutions in China. Standardized patients (SPs) presented the *same* cases of unstable angina and asthma to 248 PHC physicians and to 2 widely used AI chatbots, ChatGPT-4o and DeepSeek-R1, on the *same* day. The analysis included 744 SP encounters.

Exposures

Consultation with a PHC physician, ChatGPT-4o, or DeepSeek-R1 for the same conditions.

Main Outcomes and Measures

Primary outcomes included diagnostic accuracy (correct diagnosis and correct medication), clinical safety (number and rate of unnecessary tests and medication prescribing), and patient-centeredness (range 1–4, with higher scores indicating more patient-centered communications).

Results

Among 248 PHC physicians, 66.9% were male and the mean (SD) age was 44.7 (9.4) years. Compared with PHC physicians, ChatGPT-4o and DeepSeek-R1 had higher rates of correct diagnosis (94.0% and 99.2% vs 30.6%; P<.001), correct medication recommendation (100.0% and 100.0% vs 35.4%; P<.001), and higher patient-centeredness scores (3.28 and 3.35 vs 2.55; P<.001), particularly in *understanding the whole person* and *finding the common ground*. However, both chatbots ordered more tests and recommended more medications than physicians, resulting in higher rates of unnecessary testing (90.3% and 96.4% vs 33.1%; P<.001) and inappropriate medication recommendations (73.4% and 62.1% vs 23.4%; P<.001). Findings were robust in multivariable regression models and when local physician performance was benchmarked against a nationally representative sample of PHC providers in China.

Conclusions and Relevance

In this field experiment, patient-facing AI chatbots had higher diagnostic accuracy and patient-centered communication scores than PHC physicians, but were substantially more likely to recommend unnecessary tests and inappropriate medications. These findings suggest that patient-facing AI may improve access to diagnostic and communicative support in under-resourced primary care, while also creating a risk of persuasive, AI-induced low-value care. Governance of patient-facing AI should therefore focus not only on diagnostic accuracy, but also on preventing excessive recommendations that may shape patient expectations before clinical encounters.

Introduction

Patients seek health advice where it is easiest to obtain; increasingly, that means low-cost easily accessible patient-facing AI rather than primary care physicians. Primary care is often patients' first contact with the health system and is central to comprehensive, community-based care[1]. Strong primary care systems are associated with better population health[2] and progress toward universal health coverage[3]. Yet primary health care (PHC) worldwide faces a growing capacity crisis: rising demand from ageing populations and chronic diseases is outpacing the available workforce[4]. The WHO has projected a global shortfall of 18 million health workers by 2030, with the most severe deficits in PHC and in low- and middle-income countries (LMICs)[5]. Even in the US, the shortage of PHC physicians is projected to reach 55,200 by 2033[6]. Nearly 4.5 billion people lacked full access to essential PHC services in 2021[7]. Strengthening PHC could prevent an estimated 60 million deaths globally by 2030 and increase life expectancy by 3.7 years[7].

The PHC physician shortages contribute to long patient waiting time[8], higher clinician workload[9], sub-optimal care quality[10], and less affordable care[11]. For example, a 2026 poll reported that 36% of US adults skipped or postponed needed healthcare in the past 12 months because of cost; this rose to 75% among uninsured adults under 65[12]. In Australia, 29% of adults who needed to see a GP reported delaying or not seeing one in 2023–24, and 28% reported waiting longer than they felt acceptable[13,14]. Even when PHC services are available, quality may be undesirably low, especially in LMICs. For example, in China, only around one-quarter of diagnoses for chronic diseases and about one-third of subsequent medication prescriptions by PHC physicians were deemed accurate and appropriate according to the standard clinical guidelines[15–18]; other LMICs such as India, Ghana, Kenya, and Vietnam share the same challenge[19,20].

The emergence of AI offers promising global solutions as lab studies show that AI chatbots can match or even exceed physician performance in medical exams[21–23]. Along with AI tools being increasingly adopted in clinical practice[24,25], the general public is also turning to publicly available AI chatbots like ChatGPT for everyday health advice. For example, a 2026 poll reported that 32% of US adults had used AI chatbots for health advice to get immediate information in the past year[12]. In Australia, 9.9% of adults reported asking ChatGPT health-related questions in 2024[26]; the rate increased significantly to 64%

in early 2026[27], with higher use among groups who may face barriers to care. In China, a 2025 survey reported nearly 90% of adults believed that AI could improve the quality of primary care[28]. This growing public momentum creates an urgent need for real-world evidence on how AI chatbots perform relative to practicing PHC physicians, particularly because patients used AI-generated advice deciding whether, when, and how to consult a clinician[12]. Moreover, high diagnostic accuracy does not necessarily mean safe care; benchmark-style evaluations for AI tools may fail to capture the constraints and trade-offs of real-world clinical decision-making[29–32]. This is particularly important because patient-facing AI advice is generated outside formal health systems, while its consequences may later need to be managed within them.

To address this urgent gap, we conducted a field experiment using standardized patients (SPs), a method widely regarded as the gold standard for assessing primary care quality[33], to compare two widely used AI chatbots, ChatGPT-4o[34] and DeepSeek-R1[35], with practicing PHC physicians in China. Each SP presented the *same* scripted case to a physician and, on the *same* day, to each chatbot in a new and independent chat session. We focused on guideline-based clinical decision outputs and point-of-care safety proxies, including diagnostic accuracy, evidence-based treatment, unnecessary testing and medication, and patient-centered communication. By embedding a head-to-head comparison within real-world conditions, we provide timely evidence on how patient-facing AI may impact primary health care.

Methods

Study Setting

We conducted a field experiment in 62 PHC institutions in Henan, China during February 6–27, 2025 (Fig. 1). Henan uniquely represents China's resource-constrained healthcare landscape and a multistage cluster sampling strategy was applied: at the municipal level, *Luohe* and *Xuchang* were selected as primary sampling units; two counties were chosen from *Luohe* (*Wuyang* and *Linying*) and two from *Xuchang* (*Xiangcheng* and *Changge*). All 62 township hospitals (PHC facilities) within each county were included, comprising 14 in *Wuyang* and 16 each in *Linying*, *Xiangcheng* and *Changge*, respectively. Township hospitals represent the backbone of China's PHC system, serving as the first point of contact for most rural and peri-urban residents. Physicians in township hospitals in *Luohe* and *Xuchang* are often male (55.9%), with relatively low rates of bachelor's degree attainment (33.1%) and few holding senior titles (< 8.5%), though 21.3% hold intermediate titles[36,37].

Standardized Patients (SPs)

SPs are healthy individuals trained to consistently portray clinical cases during their visits without disclosing their role to healthcare providers. SPs follow predefined responses to physicians' questions, allowing direct observation of clinical processes and decision-making under natural conditions. SPs are considered the gold standard for measuring the quality of primary care and have been increasingly used in resource-constrained settings[38,39]. Two common conditions, unstable angina and asthma, were used based on their prevalence and availability of established management guidelines[40,41]. A panel of 6 senior doctors and pharmacists with over 15 years of clinical experience at tertiary hospitals independently reviewed and validated the scripts [15,42]. The detailed SP scripts can be found in *Supplement 1*.

To be comparable, AI chatbots were designated as a PHC physician to offer medical consultations[43]. Each SP presented a primary complaint (i.e., chest pain or shortness of breath) and then provided consistent, predefined responses to all subsequent questions from physicians or AI chatbots[15,43–45] *(*Fig. 1*)*. Each encounter was recorded individually. For each AI visit, a new, independent chat session was initiated for each encounter to avoid memory retention effects. AI chatbots' memory was cleared before each new chat. This ensured that AI chatbots treated each interaction as a separate, first-time consultation, maintaining consistency and real-world relevance of the outputs.

In the experiment, 8 SPs were rigorously trained based on script memorization, supervised rehearsals, and a pilot visit, and their performance was highly comparable before the visits. In addition, each SP was paired with an investigator, who recorded the encounters using structured questionnaires based on SP recall and medical receipts. Each SP conducted in-person visits with PHC physicians and virtual visits with AI chatbots within 15 minutes on the same day. A different investigator–SP pair repeated the same case visit at the same PHC institution on the following day. This resulted in a total of 248 physician visits across the 4 counties, paired with 496 parallel consultations using ChatGPT-4o and DeepSeek-R1. The final sample comprised 744 SP visits. Throughout the study, no PHC physician questioned the SPs' identities. Data were administered using the Research Electronic Data Capture (REDCap) system.

Outcome Indicators

Quality[46] was assessed using two continuous measures: adherence to the full checklist (range 0–1, covering all clinical inquiries and recommended tests) and adherence to the essential checklist (range 0–1, focusing on core items in the checklist); and two binary indicators: correct diagnosis (1 if any guideline-recommended diagnosis was recorded, 0 otherwise) and correct medication (1 if any guideline-recommended medication was prescribed, 0 if advice was palliative, irrelevant, unnecessary, or potentially harmful)[47]. The detailed checklists can be found in *Supplement 2*.

Safety was evaluated using four additional indicators[43]: (1) total number of tests ordered, (2) the presence of any unnecessary tests (0/1), (3) total number of drugs prescribed, and (4) the presence of any unnecessary drugs (0/1). We defined unnecessary tests as any diagnostic test that was not recommended by the disease-specific clinical guidelines, and inappropriate medications as any medication that was not guideline-recommended (including unnecessary or potentially harmful medications).

Patient-Centered Communication was assessed using the Chinese version of an internationally recognized Patient Perception of Patient-Centeredness (PPPC-CN) scale, a 21-item validated instrument[48]. The scale measures four components: (1) C1: *exploring health, disease and the illness experience*, (2) C2: *finding the common ground*, (3) C3: *understanding the whole person*, and (4) C4: *enhancing the patient-clinician relationship*. Items are rated on a 4-point scale (1 = not at all to 4 = completely), with the total average score and dimension-specific averages calculated.

Statistical Analysis

Descriptive analyses summarized quality, safety, and PPPC-CN indicators for the overall sample and by disease condition. Outcome indicators were analyzed using generalized linear models depending on the nature of outcome variables. Adjusted comparisons of AI chatbots versus PHC physicians were made in those models to derive the marginal effects (dy/dx) and the 95% confidence intervals (CIs). All models were adjusted for case, investigator, county, day of week, and PHC institutions. Two-sided P values < .05 were considered statistically significant. Analyses were conducted using Stata, version 18.0 (StataCorp).

Ethics Approval

Ethical approval

s were obtained from the ethic committees of Xi'an Jiaotong University (LLSBPJ-2026-643) and Luohe Medical College (No: LYZLL-2024012). Written consent forms were obtained from selected hospitals and physicians three months before the SPs' visits, but the physicians were unaware of the conditions to be tested.

Results

The study involved 248 PHC physicians (66.9% male; mean [SD] age, 44.7 [9.4] years). We first describe the performance of PHC physicians as compared with the two AI chatbots (Table 1). To improve internal validity, multivariable regression models were used to compare PHC physicians with ChatGPT-4o and DeepSeek-R1 on the *same* day with the *same* SPs (Table 2).

Table 1
Performance of PHC Physicians and AI Chatbots on Quality, Safety, and Patient-Centeredness in Medical

Variables	Overall (n = 744)			Unstable angina (n = 372)		
	PHC physicians	ChatGPT-4o	DeepSeek-R1	PHC physicians	ChatGPT-4o	DeepSeek-R1
Quality indicators						
Full checklist (%)	0.259(0.243,0.274)	0.416(0.407,0.425)	0.481(0.469,0.492)	0.280(0.255,0.304)	0.442(0.432,0.451)	0.537(0.527,0.546)
Essential checklists (%)	0.405(0.368,0.441)	0.554(0.526,0.582)	0.628(0.598,0.658)	0.611(0.568,0.653)	0.729(0.706,0.753)	0.840(0.824,0.855)
Correct diagnosis (0/1)	0.306(0.249,0.364)	0.940(0.910,0.969)	0.992(0.971,0.998)	0.565(0.476,0.653)	1.000(1.000,1.000)	1.000(1.000,1.000)
Correct medication (0/1)	0.354(0.234,0.473)	1.000(1.000,1.000)	1.000(1.000,1.000)	0.417(0.204,0.639)	1.000(1.000,1.000)	1.000(1.000,1.000)
Safety indicators						
No. of tests requested	2.492(2.310,2.673)	3.669(3.523,3.816)	4.891(4.689,5.093)	2.161(1.942,2.380)	3.968(3.739,4.197)	5.548(5.250,5.847)
Unnecessary test (0/1)	0.331(0.272,0.390)	0.903(0.866,0.940)	0.964(0.940,0.987)	0.282(0.202,0.363)	0.935(0.892,0.979)	0.984(0.943,0.996)
No. of medication prescribed	0.657(0.494,0.821)	5.556(5.399,5.714)	5.867(5.700,6.034)	0.323(0.184,0.461)	6.129(5.944,6.315)	6.339(6.141,6.536)
Inappropriate medication (0/1)	0.234(0.181,0.287)	0.734(0.678,0.789)	0.621(0.560,0.682)	0.153(0.089,0.218)	0.871(0.811,0.931)	0.790(0.718,0.863)
PPPC-CN						
Total score (range 1–4)	2.548(2.474,2.622)	3.281(3.218,3.343)	3.353(3.291,3.415)	2.584(2.494,2.675)	3.376(3.288,3.464)	3.423(3.333,3.513)
C1: Exploring experience	3.006(2.924,3.088)	3.505(3.450,3.561)	3.559(3.505,3.613)	3.073(2.982,3.163)	3.610(3.539,3.682)	3.638(3.568,3.709)
C2: Finding common ground	2.307(2.231,2.382)	3.116(3.044,3.188)	3.212(3.140,3.284)	2.344(2.242,2.447)	3.222(3.112,3.333)	3.286(3.173,3.399)
C3: Understanding the whole person	2.374(2.291,2.457)	3.238(3.163,3.313)	3.305(3.235,3.376)	2.349(2.260,2.437)	3.280(3.184,3.376)	3.333(3.236,3.429)
C4: Enhancing relationship	2.871(2.775,2.967)	3.499(3.440,3.558)	3.543(3.486,3.600)	2.890(2.779,3.001)	3.626(3.562,3.690)	3.642(3.576,3.709)

Note. PHC: primary health care; AI: Artificial Intelligence; PPPC-CN: Patient Perception of Patient-Centeredness. This table describes the performance of PH DeepSeek-R1) on a comprehensive set of quality, safety and patient-centered communication measures. All values are presented as means with 95% confid and 95% CIs. The sample includes 744 standardized patient visits, equally split between two disease cases: unstable angina (n = 372) and asthma (n = 372).

• Quality indicators include the adherence to the full and essential checklists (continuous variables as the percentage % of completeness) and diagnostic ac scored 0/1).

• Safety indicators capture test ordering and prescribing behavior, including the number of tests/medications (count variables) and the proportion of standa or drugs (binary variables scored 0/1).

• PPPC-CN is a patient-centeredness scale (range 1–4) with four components: Exploring health, disease and the illness experience (C1), Finding the commo the patient-clinician relationship (C4).

Table 2
Adjusted Differences (Marginal Effects) in Quality, Safety, and PPPC-CN Performance: AI Chatbots Compared with PHC Physicians and with Each Other

Variables	ChatGPT-4o (ref: PHC physicians)			DeepSeek-R1 (ref: PHC physicians)			DeepSeek-R1 (ref: ChatGPT 4o)		
	dy/dx	95% CI	*P*	dy/dx	95% CI	*P*	dy/dx	95% CI	*P*
Quality indicators									
Full checklists (%)	0.156	0.140,0.172	< 0.001	0.222	0.206,0.239	< 0.001	0.064	0.053,0.075	< 0.001
Essential checklists (%)	0.149	0.119,0.178	< 0.001	0.225	0.198,0.252	< 0.001	0.073	0.053,0.094	< 0.001
Correct diagnosis (0/1)	0.507	0.469,0.545	< 0.001	0.509	0.447,0.572	< 0.001	0.089	0.031,0.146	0.003
Correct medication (0/1)	0.646	0.527,0.766	< 0.001	0.646	0.527,0.766	< 0.001	0.000	0.000,0.000	1.000
Safety indicators									
No. of tests requested	1.183	0.951,1.414	< 0.001	2.407	2.143,2.670	< 0.001	1.218	0.988,1.448	< 0.001
Unnecessary test requested (0/1)	0.452	0.423,0.480	< 0.001	0.487	0.462,0.512	< 0.001	0.074	0.024,0.123	0.003
No. of medication prescribed	6.661	6.016,7.307	< 0.001	7.184	6.500,7.868	< 0.001	0.320	0.124,0.515	0.001
Inappropriate medication prescribed (0/1)	0.441	0.401,0.480	< 0.001	0.367	0.310,0.425	< 0.001	-0.109	-0.179,-0.038	0.003
PPPC-CN									
Total score (range 1–4)	0.722	0.659,0.785	< 0.001	0.794	0.731,0.857	< 0.001	0.072	0.039,0.105	< 0.001
C1: Exploring health, disease and the illness experience	0.489	0.429,0.549	< 0.001	0.539	0.479,0.600	< 0.001	0.053	0.022,0.085	0.001
C2: Finding the common ground	0.795	0.722,0.869	< 0.001	0.891	0.818,0.965	< 0.001	0.097	0.055,0.138	< 0.001
C3: Understanding the whole person	0.858	0.782,0.933	< 0.001	0.926	0.855,0.998	< 0.001	0.066	0.014,0.117	0.012
C4: Enhancing the patient-clinician relationship	0.625	0.544,0.706	< 0.001	0.672	0.593,0.752	< 0.001	0.040	-0.006,0.085	0.085

Note: This table presents adjusted comparisons of AI chatbots versus PHC physicians, and AI chatbots versus each other. Each coefficient (dy/dx) represents a regression model on specific outcome metrics. All models are adjusted for case, investigator, county, day of week, and PHC institutions; Marginal effects (dy/dx) and the 95% confidence intervals (CIs) were estimated by using the Generalized Linear model (Full checklist, essential checklists and PPPC-CN), Negative binomial model (No. of tests requested and No. of medication prescribed), or Probit model (Correct diagnosis, Correct medication, Unnecessary test requested, Inappropriate medication). Robust standard errors were clustered at the institution level. All P-values are two-sided. The interpretation of dy/dx depends on the outcome type:

• For binary variables (Correct diagnosis, Correct medication, Unnecessary test requested, Inappropriate medication prescribed), dy/dx is the adjusted difference in probability (range 0–1 or 0%-100%).

• For count variables (No. of tests requested, No. of medication prescribed), dy/dx is the adjusted difference in the expected count.

• For continuous variables (PPPC-CN total and its components, Full/Essential checklists as proportions), dy/dx is the adjusted difference in mean score.

Diagnostic Accuracy and Adherence to Checklist

ChatGPT-4o and DeepSeek-R1 performed better than PHC physicians on all quality measures (Table 1). Compared with physicians, ChatGPT-4o was 50.7 percentage points (pp) more likely to reach a correct diagnosis (95% CI, 46.9–54.5; P<.001) and DeepSeek-R1 was 50.9 pp more likely (95% CI, 44.7–57.2; P<.001). Both chatbots were 64.6 pp more likely to provide a correct medication (95% CI, 52.7–76.6; P<.001). DeepSeek-R1 slightly outperformed ChatGPT-4o, but the differences were small (Table 2).

The higher diagnostic accuracy of AI chatbots was partly explained by greater adherence to checklist items. Compared with physicians, ChatGPT-4o completed 15.6% more items on the full checklist (95% CI, 14.0%–17.2%; P<.001) and 14.9% more on the essential checklist (95% CI, 11.9%–17.8%; P<.001), while DeepSeek-R1 completed 22.2% more on the full checklist (95% CI, 19.8%–25.2%; P<.001) and 22.5% more on the essential checklist (95% CI, 19.8%–25.2%; P<.001). Scatterplots of checklist adherence and diagnostic accuracy demonstrated a positive association (Fig. 2), with AI chatbots having higher intercepts and relatively flat slopes compared with physicians. In addition, adherence patterns differed (*Figure S1*): AI chatbots concentrated on a narrower set of items, whereas physicians distributed their inquiries more evenly across checklist elements.

Test Ordering and Medication Prescription

Safety outcomes revealed substantially greater risks with AI chatbots (Table 1), including prescribing more tests and medications, and having a higher probability of unnecessary testing and inappropriate prescribing. Compared with physicians (Table 2), ChatGPT-4o ordered 1.18 more tests (95% CI, 0.95–1.41; P<.001) and 6.66 more medications (95% CI, 6.02–7.31; P<.001), while DeepSeek-R1 ordered 2.41 more tests (95% CI, 2.14–2.67; P<.001) and 7.18 more medications (95% CI, 6.50–7.86; P<.001). As a result, the proportion of visits involving unnecessary tests was 45.2 pp higher for ChatGPT-4o (95% CI, 42.3–48.0; P<.001) and 48.7 pp higher for DeepSeek-R1 (95% CI, 46.2–51.2; P<.001). Similarly, the proportion of visits with inappropriate medications was 44.1 pp higher for ChatGPT-4o (95% CI, 40.1–48.0; P<.001) and 36.7 pp higher for DeepSeek-R1 (95% CI, 31.0–42.5; P<.001) than for physicians.

Patient-Centered Communications

Both AI chatbots scored higher on patient-centeredness scores than PHC physicians (Table 1). Compared with PHC physicians (Table 2), total scores of PPPC-CN were 0.72 points higher for ChatGPT-4o (95% CI, 0.66–0.79; P<.001) and 0.79 points higher for DeepSeek-R1 (95% CI, 0.73–0.86; P<.001). The largest advantages were observed in *understanding the whole person* and *finding the common ground*. Communication styles also differed between physicians and AI chatbots (Table 3). In *exploring health, disease and the illness experience*, physicians most frequently emphasized whether they had identified the main reason for the visit, while AI chatbots emphasized whether the interaction conveyed compassion or empathic understanding. In *finding the common ground*, physicians most often highlighted whether the patient agreed with the physician's opinion, whereas AI chatbots emphasized whether the treatment plan had been clearly explained. See one example in *Supplement 3*.

Table 3
PPPC-CN evaluation scores by dimension for PHC physicians and AI Chatbots (N = 744)

Question Items		PHC physicians		ChatGPT-4o		DeepSeek-R1	
		Mean (95%CI)	Rank	Mean (95%CI)	Rank	Mean (95%CI)	Rank
Component 1: Exploring health, disease and the illness experience							
1	To what extent was your main problem discussed?	2.679 (2.579, 2.780)	5	3.512 (3.444, 3.589)	4	3.547 (3.481, 3.613)	5
2	Does the doctor know that this is the main reason you came to the clinic today?	3.385 (3.282, 3.487)	1	3.659 (3.591, 3.727)	2	3.673 (3.611, 3.736)	2
3	To what extent did the doctor understand the importance of the problem?	3.269 (3.159, 3.379)	2	3.597 (3.527, 3.667)	3	3.612 (3.546, 3.679)	3
4	How well do you think the doctor understood your words during the consultation?	2.846 (2.752, 2.939)	4	3.479 (3.413, 3.546)	5	3.592 (3.527, 3.658)	4
19*	To what extent does the doctor show you compassion (or empathic understanding)?	3.141 (3.035, 3.247)	3	3.686 (3.627, 3.745)	1	3.722 (3.665, 3.780)	1
14*	How much would you say that the doctor pays comprehensive attention to you?	2.471 (2.369, 2.571)	6	3.066 (2.956, 3.175)	6	3.192 (3.087, 3.296)	6
Component 2: Finding the common ground							
5	How satisfied were you with the discussion of your problem?	2.316 (2.198, 2.434)	6	3.331 (3.263, 3.398)	5	3.490 (3.422, 3.558)	3
6	To what extent did the doctor explain this problem to you?	2.573 (2.490, 2.655)	1	3.450 (3.386, 3.515)	1	3.478 (3.415, 3.541)	1
7	To what extent did you agree with the doctor's opinion about your problem?	2.568 (2.487, 2.649)	2	3.359 (3.295, 3.423)	4	3.449 (3.386, 3.512)	5
8	How many opportunities did you have to ask questions?	2.474 (2.379, 2.569)	3	2.969 (2.837, 3.096)	7	3.037 (2.910, 3.164)	8
9	To what extent did the doctor ask about your goals for treatment?	2.299 (2.204, 2.394)	7	3.364 (3.288, 3.439)	3	3.465 (3.395, 3.536)	4
10	To what extent did the doctor explain your treatment?	2.337 (2.232, 2.442)	5	3.450 (3.384, 3.516)	2	3.498 (3.435, 3.561)	2
11	To what extent does the doctor discuss the feasibility of this treatment with you?	1.962 (1.869, 2.055)	8	2.950 (2.858, 3.043)	8	3.074 (2.978, 3.170)	7
12	To what extent did you and your doctor discuss your respective roles?#	1.650 (1.551, 1.747)	10	2.475 (2.340, 2.610)	10	2.624 (2.483, 2.766)	10
13	To what extent did the doctor encourage you to take your role you in the care process?	1.808 (1.694, 1.921)	9	2.669 (2.537, 2.802)	9	2.784 (2.652, 2.915)	9
14*	To what extent did the doctor consider your thoughts and feelings?	2.471 (2.369, 2.571)	4	3.066 (2.956, 3.175)	6	3.192 (3.087, 3.296)	6
Component 3: Understanding the whole person							
15	How much would you say that the doctor pays comprehensive attention to you?	2.645 (2.537, 2.753)	2	3.467 (3.385, 3.549)	2	3.522 (3.452, 3.593)	2
16	To what extent does the doctor know your family circumstances about your condition?&	1.239 (1.162, 1.316)	4	2.777 (2.651, 2.903)	4	2.931 (2.813, 3.048)	4
17	Do you feel comfortable discussing your personal health issues with your doctor?	2.953 (2.828, 3.077)	1	3.634 (3.571, 3.702)	1	3.653 (3.592, 3.712)	1
18	To what extent did your doctor respect your values, customs, or personal perception?	2.504 (2.389, 2.618)	3	3.041 (2.930, 3.152)	3	3.102 (2.992, 3.212)	3
Component 4: Enhancing the patient-clinician relationship							
19*	To what extent does the doctor show you compassion (or empathic understanding)?	3.141 (3.035, 3.247)	1	3.686 (3.627, 3.745)	1	3.722 (3.665, 3.780)	1
20	To what extent does the doctor carefully listen to what you want to say?	2.654 (2.540, 2.767)	3	3.256 (3.150, 3.362)	3	3.269 (3.167, 3.372)	3
21	To what extent do you trust your doctor?	2.675 (2.566, 2.784)	2	3.570 (3.506, 3.634)	2	3.649 (3.588, 3.710)	2

Note: * Across two dimensions; # namely, who is responsible for making decisions and for certain aspects of your care; & e.g., family history of illness, family circumstances that affect recovery, etc.

Sensitivity analysis and external validity of the experiment

We conducted two sensitivity analyses. First, AI chatbots may gain an advantage by offering multiple diagnoses or treatments in a single encounter. Diagnostic accuracy was re-evaluated under two scenarios: (1) any correct, in which accuracy was credited if any diagnosis or medication matched guidelines, and (2) first correct, in which only the first-listed diagnosis or medication was considered. When restricted to the first-listed diagnosis (*Table S1*), accuracy for physicians remained 30.6%, while it declined to 76.2% for ChatGPT-4o and 87.5% for DeepSeek-R1. A similar pattern was observed for medications. Second, we conducted a reverse-ordering sensitivity analysis. On each study day, one SP team visited four institutions sequentially, completing a physician visit followed by two AI consultations at each. By dropping the first physician visit (the worst SP performance if there was an order effect) and the last pair of AI consultations within each day (the best SP performance), we re-matched the remaining observations such that all AI consultations temporally preceded the paired physician visit. Results were robust to this re-ordering (*Table S2*).

To assess external validity, we benchmarked local physician performance against the ACACIA (Primary heAlth Care quAlity Cohort In ChinA) dataset, which constitutes a nationally representative sample of primary care providers from seven geographically diverse provinces in China[49,50]. Overall, Henan physicians were broadly comparable with the national sample across all domains. The details were reported in *Table S3*.

Discussion

Low-cost easily accessible AI chatbots are increasingly used by patients to decide whether, when, and how to consult a clinician[12], creating an urgent demand for empirical evidence to inform policy and practice. In this field experiment in China, two widely used AI chatbots, ChatGPT-4o and DeepSeek-R1, showed higher diagnostic accuracy and patient-centered communication than practicing PHC physicians but also recommended substantially more low-value care. These findings reveal a central trade-off: patient-facing AI has the potential to augment the diagnostic and communicative support of primary care in under-resourced settings but also raise serious concerns about clinical safety. Without appropriate regulation and governance, patients may enter consultations with persuasive AI advice for medical tests or medications that are not clinically indicated, placing additional pressure on PHC physicians to redirect care.

ChatGPT-4o and DeepSeek-R1 achieved very high diagnostic accuracy (94.0% and 99.2%), compared with 30.6% among PHC physicians in China. These findings move beyond laboratory and benchmark-style evaluations to real-world primary care[21,43,51], where clinical information is elicited through interaction and decisions are shaped by uncertainty. The relevance of this finding also extends far beyond: physician performance in our sample was broadly comparable with a national primary care sample; globally previous SP studies in other settings, including India and Kenya, have reported correct diagnosis rates of only 12–52% in primary care[15,20]. The higher diagnostic accuracy of AI chatbots was partly driven by a greater adherence to clinical checklists as physicians often do less than what they know they should do in practice (*the know–do gap*)[52]. The results suggest that AI chatbots could augment the diagnostic capacity of PHC in resource-constrained health systems, particularly where missed or delayed diagnosis remains common.

Despite the high diagnostic accuracy, our findings highlight important safety concerns about patient-facing AI. ChatGPT-4o and DeepSeek-R1 requested 47–96% more tests and recommended more than eight times as many medications as physicians; consequently, more than 90% of AI encounters involved at least one unnecessary test. This tendency toward over-recommendation may not be limited to a single model as earlier pilot studies using ChatGPT-3.5[45] and ERNIE Bot 3.5[43] suggested similar patterns. The persistent patterns of patient-facing AI suggest a risk-averse and exhaustive strategy, optimized for completeness and omission avoidance rather than the selective clinical judgement expected in healthcare. A key structural driver of the problem may be the alignment strategies during training, including reinforcement learning from human feedback, which often reward answers that appear comprehensive, cautious and helpful, consistent with reward-induced sycophancy documented in other high-stakes AI contexts[53]. Such “zero-risk” AI behavior creates substantial safety concerns for patients and inefficiency for health systems.

AI chatbots produced communications that SPs perceived as more patient-centered, despite common concerns that AI lacks genuine empathy and human connection[54]. Emerging evidence suggested that AI agents can be acceptable to patients without time pressure[55,56] and judgmental tone[57,58]. In the study, the largest gaps were observed in *understanding the whole person* and *finding the common ground*, suggesting that the chatbots’ communication advantage was driven by their tendency to explicitly acknowledge the patient’s circumstances, and present treatment plans in ways that appeared more understandable and collaborative. However, highly patient-centered AI communication could also make its low-value recommendations more persuasive. In settings where physicians often face structural time and resource constraints[17] and patients already face substantial information deficits, patient-facing AI chatbots that communicate warmly, confidently, and in detail may generate or amplify demand for unnecessary tests or medications.

By now, most patients have not yet received guidance on the safe and appropriate use of AI tools in health contexts[26], making them particularly vulnerable to over-reliance on AI advice, particularly when patient-facing AI provides recommendations that are warm, detailed, confident, and clinically excessive. In resource-constrained PHC settings, physicians may not have the time, communication capacity, or institutional support to reverse these expectations shaped before the consultation. A recent US case has already illustrated how AI advice can influence patient behavior, leaving clinicians to manage consequences rather than preventing unsafe recommendations at the point of generation[59]. Moreover, our use of neutral prompts approximated how the public may ordinarily access these tools but AI behavior is highly context-dependent and malleable[60,61]. Real-world outputs may therefore be more complex than those documented in the study. Future research should examine whether prompt strategies or system-level instructions can preserve the diagnostic advantages of AI chatbots while reducing low-value testing and prescribing.

This study has limitations. First, we used two clinical scenarios, which might not capture the full spectrum of primary care presentations. However, the presenting complaints (e.g., shortness of breath) are common, enhancing the relevance and practical value of this study. Previous studies have also

validated that SPs closely mirror behaviors of actual patients[19,62]. Finally, the experiment was conducted in one province in China and assessed only single encounters between SPs and providers, although physician performance in Henan was broadly comparable with national data. The long-term effects of AI use on physician practice, care continuity, physician–patient relationships, and patient outcomes therefore remain unknown.

In conclusion, in this field experiment, two widely used patient-facing AI chatbots achieved substantially higher diagnostic accuracy and patient-centered communication than practicing PHC physicians, but also recommended more tests and medications, including many not indicated by guidelines. These findings suggest that patient-facing AI may help address diagnostic and communication gaps in under-resourced primary care, while also amplifying low-value care through persuasive communication. For health systems, the central question is therefore not only whether AI can provide correct answers, but whether patient-facing AI can be governed so that it improves care-seeking without creating unsafe expectations that frontline physicians are left to manage.

Declarations

Conflict of Interest

The authors report no conflicts of interest.

Funding

This work was supported by the leading talents project in philosophy and social sciences of the National Social Science Foundation of China (2022LJRC02), National Natural Science Foundation of China (72204128, 72164031), Scientific Research Program for Young Education Innovation Teams in Shaanxi Province (24JP100), Xi'an Jiaotong University Key Project under the 'National Priorities' Special Research Program, Ministry of Education Youth Fund Project for Humanities and Social Sciences Research, the Drazen Scholarship and the Aden Scholarship dedicated to research on Chinese healthcare systems. The funders had no role in the design and conduct of the study; collection, management, analysis, or interpretation of the data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication.

Data Availability

The physician data for this study is not openly available as the participants of this study did not give consent for their data to be shared publicly. Data from the SP simulations are available with publication. The data are available to anyone who requests them for any non-commercial purposes. The data can be accessed by contacting YS (yafei.si@unimelb.edu.au), who will provide guidance on how to use and interpret the data.

Data Access and Responsibility

YS, SG, ZZ and XF had full access to all the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis.

Author Contributions

YS and ZZ conceived the study; YS, XF, and ZZ designed the study with the advisory committee of GC, XC, LM, BL and HB; XF and SG collected the data; XF and YS drafted the main manuscript text; XF and YS prepared tables and figures. MT, HB, BL, SC, DW, RA, JZ, MS, YM, SA, HZ, SG, ZZ, LM, XC, and GC edited the manuscript. All authors reviewed and approved the final manuscript.

Acknowledgement

ChatGPT-5.5, OpenAI, was used in May–June 2026 to assist with language editing and to improve the clarity and readability of the manuscript. All AI-assisted outputs were critically reviewed and edited by the authors, who remain fully responsible for the accuracy, integrity, and final content of the manuscript.

References

1. Beaudevin C, Gaudillière JP, Gradmann C. The local roots of 'health for all': Primary health care in practices, 1950s–2000s. *Soc Sci Med*. 2023;319:115321. doi:10.1016/j.socscimed.2022.115321
2. Macinko J, Starfield B, Shi L. The Contribution of Primary Care Systems to Health Outcomes within Organization for Economic Cooperation and Development (OECD) Countries, 1970–1998. *Health Serv Res*. 2003;38(3):831-865. doi:10.1111/1475-6773.00149
3. Hone T, Macinko J, Millett C. Revisiting Alma-Ata: what is the role of primary health care in achieving the Sustainable Development Goals? *The Lancet*. 2018;392(10156):1461-1472. doi:10.1016/S0140-6736(18)31829-4
4. Russo G, Perelman J, Zapata T, Šantrić-Milićević M. The layered crisis of the primary care medical workforce in the European region: what evidence do we need to identify causes and solutions? *Hum Resour Health*. 2023;21(1):55, s12960-023-00842-00844. doi:10.1186/s12960-023-00842-4
5. Haakenstad A, Irvine CMS, Knight M, et al. Measuring the availability of human resources for health and its relationship to universal health coverage for 204 countries and territories from 1990 to 2019: a systematic analysis for the Global Burden of Disease Study 2019. *The Lancet*. 2022;399(10341):2129-2154. doi:10.1016/S0140-6736(22)00532-3
6. Association of American Medical Colleges. The complexities of physician supply and demand: projections from 2018 to 2033. Published online November 27, 2020. https://www.aamc.org/media/45976/download
7. WHO. Universal health coverage (UHC). Published online March 26, 2025. https://www.who.int/news-room/fact-sheets/detail/universal-health-coverage-(uhc)

8. Feyman Y, Asfaw DA, Griffith KN. Geographic variation in appointment wait times for US military veterans. *JAMA Netw Open*. 2022;5(8):e2228783.
9. UNICEF. Investing in health workers to save maternal and newborn lives. *Brief Note*. Published online 2020.
10. Albarqouni L, Palagama S, Chai J, et al. Overuse of medications in low- and middle-income countries: a scoping review. *Bull World Health Organ*. 2023;101(1):36-61D. doi:10.2471/BLT.22.288293
11. Albarqouni L, Arab-Zozani M, Abukmail E, et al. Overdiagnosis and overuse of diagnostic and screening tests in low-income and middle-income countries: a scoping review. *BMJ Glob Health*. 2022;7(10):e008696. doi:10.1136/bmjgh-2022-008696
12. kffjulianm. KFF Tracking Poll on Health Information and Trust: Use of AI For Health Information and Advice. KFF. March 25, 2026. Accessed May 7, 2026. https://www.kff.org/public-opinion/kff-tracking-poll-on-health-information-and-trust-use-of-ai-for-health-information-and-advice/
13. General practice, allied health and other primary care services. Australian Institute of Health and Welfare. March 26, 2026. Accessed April 18, 2026. https://www.aihw.gov.au/reports/primary-health-care/general-practice-allied-health-primary-care?ref=blog.halaxy.com
14. ABS. Patient Experiences, 2023-24 financial year. November 18, 2024. Accessed April 18, 2026. https://www.abs.gov.au/statistics/health/health-services/patient-experiences/2023-24
15. Si Y, Bateman H, Chen S, et al. Quantifying the financial impact of overuse in primary care in China: A standardised patient study. *Soc Sci Med*. Published online 2023:115670.
16. Si Y, Xue H, Liao H, et al. The quality of telemedicine consultations for sexually transmitted infections in China. *Health Policy Plan*. 2024;39(3):307-317.
17. Su M, Zhou Z, Si Y, Fan X. The Association Between Patient-Centered Communication and Primary Care Quality in Urban China: Evidence From a Standardized Patient Study. *Front Public Health*. 2022;9:779293.
18. Sylvia S, Shi Y, Xue H, et al. Survey using incognito standardized patients shows poor quality care in China's rural clinics. *Health Policy Plan*. 2015;30(3):322-333.
19. Das J, Chowdhury A, Hussam R, Banerjee AV. The impact of training informal health care providers in India: A randomized controlled trial. *Science*. 2016;354(6308).
20. Kwan A, Daniels B, Bergkvist S, Das V, Pai M, Das J. Use of standardised patients for healthcare quality research in low-and middle-income countries. *BMJ Glob Health*. 2019;4(5):e001669.
21. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*. 2017;542(7639):115-118. doi:10.1038/nature21056
22. McKinney SM, Sieniek M, Godbole V, et al. International evaluation of an AI system for breast cancer screening. *Nature*. 2020;577(7788):89-94. doi:10.1038/s41586-019-1799-6
23. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. *Nature*. 2023;620(7972):172-180.
24. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med*. 2019;25(1):44-56. doi:10.1038/s41591-018-0300-7
25. Mikhail D, Farah A, Milad J, et al. DeepSeek-R1 vs OpenAI o1 for Ophthalmic Diagnoses and Management Plans. *JAMA Ophthalmol*. Published online September 4, 2025. doi:10.1001/jamaophthalmol.2025.2918
26. Ayre J, Cvejic E, McCaffery KJ. Use of ChatGPT to obtain health information in Australia, 2024: insights from a nationally representative survey. *Med J Aust*. 2025;222(4):210-212.
27. TMRW. AI in Australian Healthcare: What 1,043 Adults Think About AI, Trust, and Clinically-Governed Health Services. Published online February 11, 2026. Accessed May 4, 2026. https://www.startmytomorrow.com/you-gov
28. The 2025 China Future Health Index report. *The 2025 China Future Health Index Report*. 2025.
29. Tordjman M, Liu Z, Yuce M, et al. Comparative benchmarking of the DeepSeek large language model on medical tasks and clinical reasoning. *Nat Med*. Published online 2025:1-1.
30. Kaiser KN, Hughes AJ, Yang AD, et al. Accuracy and consistency of publicly available Large Language Models as clinical decision support tools for the management of colon cancer. *J Surg Oncol*. 2024;130(5):1104-1110. doi:10.1002/jso.27821
31. Sandmann S, Hegselmann S, Fujarski M, et al. Benchmark evaluation of DeepSeek large language models in clinical decision-making. *Nat Med*. Published online 2025:1-1.
32. Tu T, Schaekermann M, Palepu A, et al. Towards conversational diagnostic artificial intelligence. *Nature*. Published online 2025:1-9.
33. Peabody JW, Luck J, Glassman P, Dresselhaus TR, Lee M. Comparison of vignettes, standardized patients, and chart abstraction: a prospective validation study of 3 methods for measuring quality. *Jama*. 2000;283(13):1715-1722.
34. Introducing GPT-4o and more tools to ChatGPT free users. Accessed October 3, 2025. https://openai.com/index/gpt-4o-and-more-tools-to-chatgpt-free/
35. Guo D, Yang D, Zhang H, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*. 2025;645(8081):633-638.
36. Henan Statistical Yearbook 2024. *Henan Statistical Yearbook 2024*. 2025.
37. National Health Commissio. *2023 China Health Statistics Yearbook*. 2024.
38. Xie Y, He W, Wan Y, et al. Validity of patients' online reviews at direct-to-consumer teleconsultation platforms: a protocol for a cross-sectional study using unannounced standardised patients. *BMJ Open*. 2023;13(5):e071783. doi:10.1136/bmjopen-2023-071783
39. Das V, Daniels B, Kwan A, et al. Simulated patients and their reality: An inquiry into theory and method. *Soc Sci Med*. 2022;300:114571. doi:10.1016/j.socscimed.2021.114571
40. Chinese Medical Association, Chinese Medical Association Journal, Chinese Medical Association General Practice Branch, Chinese Medical Association Respiratory Disease Branch Asthma Working Group, Chinese Medical Association Chinese General Practitioner Journal Editorial Committee, Expert

Group for the Compilation of Guidelines for the Diagnosis and Treatment of Respiratory Diseases at the Primary Care Level. Guideline for primary care of bronchial asthma (2018). *Chin J Gen Pract*. 2018;17(10):751-762. doi:10.3760/cma.j.issn.1671-7368.2018.10.002

41. Chinese Medical Association Cardiovascular Disease Branch, Editorial Board of Chinese Journal of Cardiovascular Disease. Guidelines for the management of non-ST elevation acute coronary syndromes (2024). *Chin J Cardiol*. 2024;52(6):615-646. doi:10.3760/cma.j.cn112148-20240107-00014
42. Su M, Zhou Z, Si Y, et al. Comparing the Quality of Primary Care between Public and Private Providers in Urban China: A Standardized Patient Study. *Int J Environ Res Public Health*. 2021;18(10):5060.
43. Si Y, Meng Y, Chen X, et al. Quality safety and disparity of an AI chatbot in managing chronic diseases: simulated patient experiments. *Npj Digit Med*. 2025;8(1):574. doi:10.1038/s41746-025-01956-w
44. Si Y, Zhou Z, Su M, Chen X. Revisiting gender gap in quality of health care in urban China: a standardised patient audit study. *The Lancet*. 2019;394:S25.
45. Si Y, Yang Y, Wang X, et al. Quality and Accountability of ChatGPT in Health Care in Low-and Middle-Income Countries: Simulated Patient Study. *J Med Internet Res*. 2024;26:e56121.
46. Das J, Woskie L, Rajbhandari R, Abbasi K, Jha A. Rethinking assumptions about delivery of healthcare: implications for universal health coverage. *BMJ*. Published online May 21, 2018:k1716. doi:10.1136/bmj.k1716
47. Si Y, Chen G, Zhou Z, Yip W, Chen X. The impact of physician-patient gender match on healthcare quality: An experiment in China. *Soc Sci Med*. 2025;380:118166. doi:10.1016/j.socscimed.2025.118166
48. Gong W, Liu L, Li X, et al. Quality of asynchronous webchats vs in-person consultations for postpartum depression in China: a cross-sectional, mixed methods study using standardized patients. *Lancet Reg Heal Pac*. 2024;45. Accessed June 2, 2026. https://www.thelancet.com/journals/lanwpc/article/PIIS2666-6065(24)00047-6/fulltext
49. Xu D (Roman), Cai Y, Wang X, et al. Improving Data Surveillance Resilience Beyond COVID-19: Experiences of Primary heAlth Care quAlity Cohort In ChinA (ACACIA) Using Unannounced Standardized Patients. *Am J Public Health*. 2022;112(6):913-922. doi:10.2105/AJPH.2022.306779
50. Zhang L, Liang H, Luo H, et al. Quality in screening and measuring blood pressure in China's primary health care: a national cross-sectional study using unannounced standardized patients. *Lancet Reg Health - West Pac*. 2024;43:100973. doi:10.1016/j.lanwpc.2023.100973
51. De Fauw J, Ledsam JR, Romera-Paredes B, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nat Med*. 2018;24(9):1342-1350. doi:10.1038/s41591-018-0107-6
52. Mohanan M, Vera-Hernández M, Das V, et al. The know-do gap in quality of health care for childhood diarrhea and pneumonia in rural India. *JAMA Pediatr*. 2015;169(4):349-357.
53. Denison C, MacDiarmid M, Barez F, et al. Sycophancy to Subterfuge: Investigating Reward-Tampering in Large Language Models. *arXiv*. Preprint posted online June 29, 2024:arXiv:2406.10162. doi:10.48550/arXiv.2406.10162
54. Vannacci A, Bonaiuti R, Ravaldi C. Machine-Made Empathy? Why Medicine Still Needs Humans. *JAMA Intern Med*. 2023;183(11):1279-1279.
55. Ayers JW, Poliak A, Dredze M, 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):589-596.
56. Ben-Haim G, Livne A, Manor U, et al. Evaluating empathy in GPT-4-generated vs. physician-written emergency department discharge letters. *Digit Health*. 2025;11:20552076251389992. doi:10.1177/20552076251389992
57. Nadarzynski T, Miles O, Cowie A, Ridge D. Acceptability of artificial intelligence (AI)-led chatbot services in healthcare: A mixed-methods study. *Digit Health*. 2019;5:2055207619871808. doi:10.1177/2055207619871808
58. Reinhardt A, Matthes J, Bojic L, Maindal HT, Paraschiv C, Ryom K. Help me, Doctor AI? A cross-national experiment on the effects of disease threat and stigma on AI health information-seeking intentions. *Comput Hum Behav*. Published online 2025:108718.
59. Eichenberger A, Thielke S, Van Buskirk A. A Case of Bromism Influenced by Use of Artificial Intelligence. *Ann Intern Med Clin Cases*. 2025;4(8):e241260. doi:10.7326/aimcc.2024.1260
60. Dorotic M, Stagno E, Warlop L. AI on the street: Context-dependent responses to artificial intelligence. *Int J Res Mark*. 2024;41(1):113-137.
61. Lu H, Lin Y, Li Z, Yiu ML, Gao Y, Uddin S. Toward fair medical advice: Addressing and mitigating bias in large language model-based healthcare applications. *Artif Intell Med*. Published online 2025:103216.
62. Das J, Holla A, Mohpal A, Muralidharan K. Quality and accountability in health care delivery: audit-study evidence from primary care in India. *Am Econ Rev*. 2016;106(12):3765-3799.

Figures

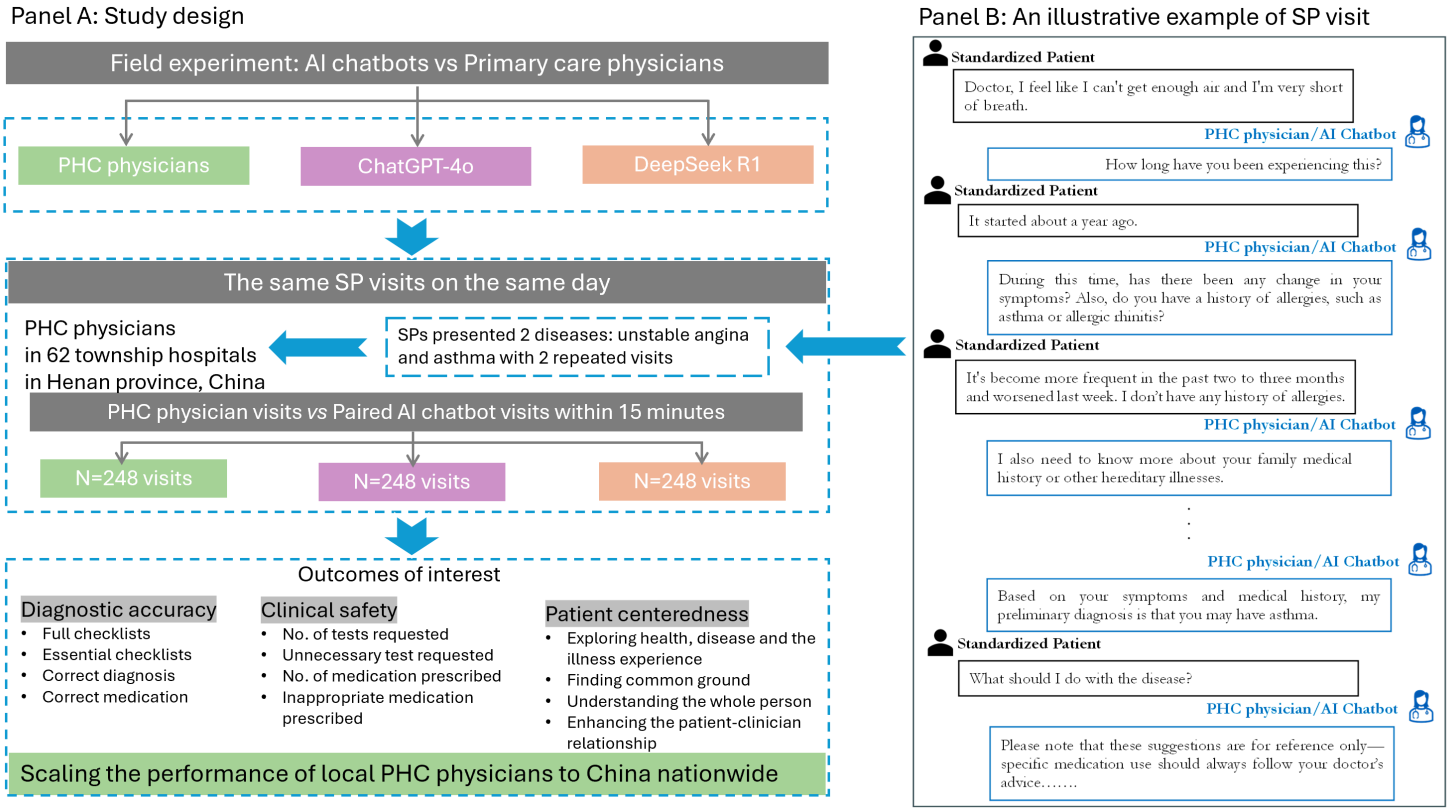


Figure 1

Flow chart of the study design

Note: For comparability, both AI chatbots were prompted to act as a Chinese primary care physician: *"You are a primary care physician in China. Please conduct routine consultations with patients."* This neutral role prompt was used to standardize chatbot consultations and approximate low-instruction public use while maintaining comparability with primary care physicians.

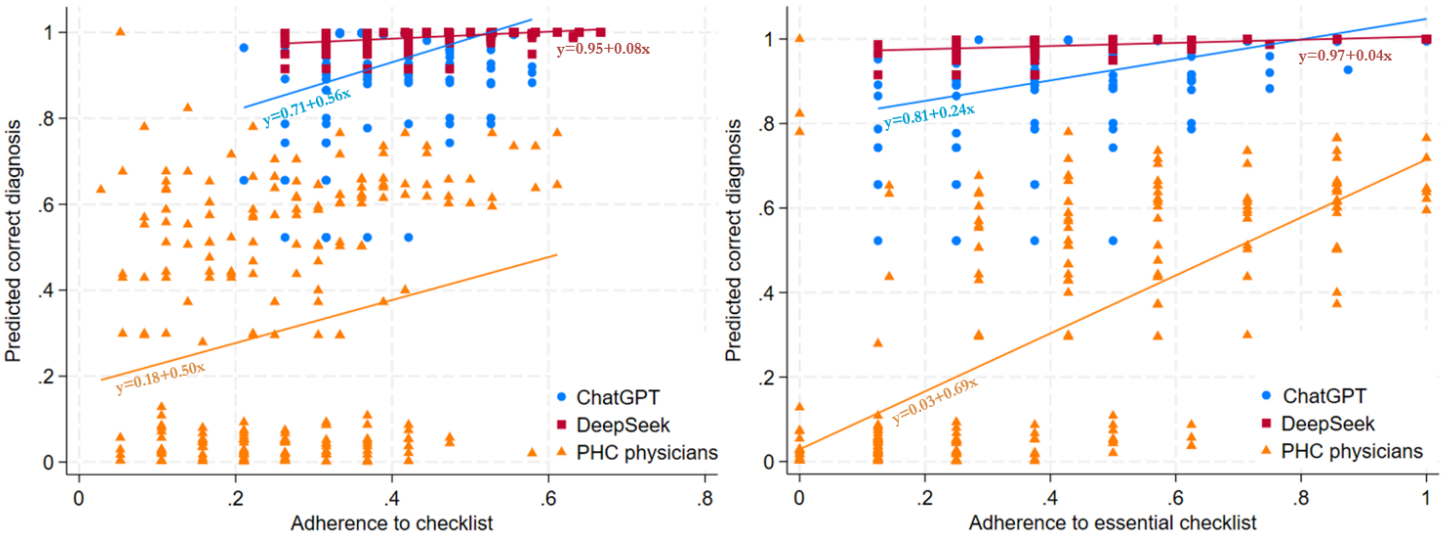


Figure 2

Relationship of adherence to checklist and correct diagnosis

Note: The linear relationship between adherence to checklist and the predicted likelihood of correct diagnosis was fitted. The associated functions are presented for physicians, ChatGPT-4o and DeepSeek-R1.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementChinaAIPhysicianExperiment26.06.10.docx