Performance Comparison of Human Doctors and Large Language Models in Tuberculosis Triage, Diagnosis, and Management:An Experimental Study

Jin Liao, Wenjun He, Huiyi Pan, Lanping Zhang, Xingyan Li, Jiamin Huang, zhichao Liu, Xue Ke, Jian Li, Xue Li, Candice Hwang, Haiting Cai, Guobao Li, jinghui Chang

Submitted to: Journal of Medical Internet Research
on: October 10, 2025

Table of Contents

Performance Comparison of Human Doctors and Large Language Models in Tuberculosis Triage, Diagnosis, and Management:An Experimental Study

Jin Liao[1*] MA (in progress); Wenjun He[2*] PhD; Huiyi Pan[1] MA(in progress); Lanping Zhang[3] MA; Xingyan Li[4] BA(in progress); Jiamin Huang[3] MS; zhichao Liu[3] BS; Xue Ke[3] MS; Jian Li[3] BS; Xue Li[3] MS; Candice Hwang[5] PhD; Haiting Cai[1] BA; Guobao Li[3] PhD; jinghui Chang[1] PhD

[1] School of Health Management, Southern Medical University, Guangzhou, China Guang Zhou CN
[2] Dermatology Hospital, Southern Medical University, Guangzhou, China Guang Zhou CN
[3] Department of the Third Pulmonary Disease, Shenzhen Third People's Hospital?Shenzhen, Guangdong Province, 518112, China Shen Zhen CN
[4] School of Health Policy and Management,Nanjing Medical University Nan Jing CN
[5] Department of Internal Medicine, Stanford University School of Medicine, Stanford, California, USA California AM
[*]these authors contributed equally

Corresponding Author:
jinghui Chang PhD

School of Health Management, Southern Medical University, Guangzhou, China
No. 1023-1063, Shatai South Road, Baiyun District
Guang Zhou
CN

Abstract

Background: Tuberculosis (TB) remains a major global health challenge, particularly in low- and middle-income countries, where effective triage, diagnosis, and management are often limited. Existing decision-support tools focus on imaging and cannot integrate multi-modal clinical information, constraining their utility in complex clinical scenarios. Large Language Models (LLMs) have shown promise in assisting diagnosis and clinical decision-making in other medical fields, but evidence for their application in TB care is scarce. Evaluating LLMs for TB decision support is crucial to explore their potential to improve clinical accuracy, efficiency, and quality of care in high-burden, resource-limited settings.

Objective: To evaluate whether large language models (LLMs) can assist tuberculosis (TB) physicians in clinical decision-making across triage, differential diagnosis, and management recommendation tasks, addressing potential delays and inequities in TB care.

Methods: In this experimental comparative study conducted in 2025 under STARD guidelines, 17 standardized TB cases (7 simulated, 10 real) were assessed. Responses were generated by two advanced LLMs (ChatGPT-4o and DeepSeek-R1) and two TB physicians. Reference standards were established by three TB specialists. Objective performance was measured using precision, recall, and F1 scores. Subjective evaluation assessed suitability, information quality, and, for management tasks, safety, conciseness, understandability, and operability using 5-point Likert scales. Readability was measured by a Chinese R-value; group differences were analyzed using Mann-Whitney U tests.

Results: LLMs achieved precision similar to physicians across all tasks (median 0.67 vs 0.50; U = 8695.5; P = .35) but higher recall (0.53 vs 0.33; U = 6848.5; P < .001) and F1 scores (0.58 vs 0.33; U = 7085.5; P < .001) in management recommendation tasks. In management tasks, LLMs outperformed physicians in recall (0.50 vs 0.20; U = 185.0; P < .001) and F1 (0.50 vs 0.30; U = 104.0; P < .001), with no difference in precision. Subjectively, LLMs scored higher in suitability (3.67 vs 3.00; U = 1122.0; P < .001), information quality (3.33 vs 2.67; U = 155.0; P < .001), understandability (3.67 vs 3.00; U = 4281.5; P = .022), and operability (3.67 vs 3.00; U = 4305.0; P = .025). No differences were observed in conciseness (P = .54) or safety (P = .06). Physicians' responses were more readable (1.88 vs 2.17; U = 11427.5; P < .001).

Conclusions: LLMs can serve as adjuncts to support TB clinical decision-making, enhancing management recommendations without replacing physicians. Their use may improve decision efficiency and help reduce disparities in TB care. Clinical Trial:

This experimental comparative study evaluating large language models versus tuberculosis physicians did not involve patient interventions or randomization, and therefore was not registered as a clinical trial.

(JMIR Preprints 10/10/2025:85613)
DOI: https://doi.org/10.2196/preprints.85613

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http
No. Please do not make my accepted manuscript PDF available to anyone. I understand that if I later pay to participate in <a href="https:/

Original Manuscript

Performance Comparison of Human Doctors and Large Language Models in Tuberculosis Triage, Diagnosis, and Management:An Experimental Study

Jin Liao[1#]，Wenjun He[2#]，Huiyi Pan[1]，Lanping Zhang[3]，Xingyan Li[4]，Jiamin Huang[3]，Zhichao Liu[3]，Xue Ke[3]，Jian Li[3]，Xue Li[3]，Candice Hwang[5] ,Haiting Cai [1], Guobao Li[3*]，Jinghui Chang[1*]

1.School of Health Management, Southern Medical University, Guangzhou, China
2.Dermatology Hospital, Southern Medical University, Guangzhou, China
3.Department of the Third Pulmonary Disease, Shenzhen Third People's Hospital，Shenzhen, Guangdong Province, China
4.School of Health Policy and Management,Nanjing Medical University，Nanjing，China
5.Department of Internal Medicine, Stanford University School of Medicine, Stanford, California, USA

Corresponding author:
Jinghui Chang, School of Health Management, Southern Medical University, Guangzhou, China
Email: erin2007@foxmail.com

Guobao Li, Department of the Third Pulmonary Disease, Shenzhen Third People's Hospital，Shenzhen, Guangdong Province, China
Email: ligb2020@mail.sustech.edu.cn

Abstract

Background:

Tuberculosis (TB) remains a major global health challenge, particularly in low- and middle-income countries, where effective triage, diagnosis, and management are often limited. Existing decision-support tools focus on imaging and cannot integrate multi-modal clinical information, constraining their utility in complex clinical scenarios. Large Language Models (LLMs) have shown promise in assisting diagnosis and clinical decision-making in other medical fields, but evidence for their application in TB care is scarce. Evaluating LLMs for TB decision support is crucial to explore their potential to improve clinical accuracy, efficiency, and quality of care in high-burden, resource-limited settings.

Objective:

To evaluate whether large language models (LLMs) can assist tuberculosis (TB) physicians in clinical decision-making across triage, differential diagnosis, and management recommendation tasks, addressing potential delays and inequities in TB care.

Methods:

In this experimental comparative study conducted in 2025 under STARD guidelines, 17 standardized TB cases (7 simulated, 10 real) were assessed. Responses were generated by two advanced LLMs (ChatGPT-4o and DeepSeek-R1) and two TB physicians. Reference standards were established by three TB specialists. Objective performance was measured using precision, recall, and F1 scores. Subjective evaluation assessed suitability, information quality, and, for management tasks, safety, conciseness, understandability, and operability using 5-point Likert scales. Readability was measured by a Chinese R-value; group differences were analyzed using Mann-Whitney U tests.

Results:

LLMs achieved precision similar to physicians across all tasks (median 0.67 vs 0.50; U = 8695.5; *P* = .35) but higher recall (0.53 vs 0.33; U = 6848.5; *P* < .001) and F1 scores (0.58 vs 0.33; U = 7085.5; *P* < .001) in management recommendation tasks. In management tasks, LLMs outperformed physicians in recall (0.50 vs 0.20; U = 185.0; *P* < .001) and F1 (0.50 vs 0.30; U = 104.0; *P* < .001), with no difference in precision. Subjectively, LLMs scored higher in suitability (3.67 vs 3.00; U = 1122.0; *P* < .001), information quality (3.33 vs 2.67; U = 155.0; *P* < .001), understandability (3.67 vs 3.00; U = 4281.5; *P* = .022), and operability (3.67 vs 3.00; U = 4305.0; *P* = .025). No differences were observed in conciseness (*P* = .54) or safety (*P* = .06). Physicians' responses were more readable (1.88 vs 2.17; U = 11427.5; *P* < .001).

Conclusion:

LLMs can serve as adjuncts to support TB clinical decision-making, enhancing management recommendations without replacing physicians. Their use may improve decision efficiency and help reduce disparities in TB care.

Keywords: Tuberculosis; Large Language Models; Clinical Decision Support

Introduction

Tuberculosis (TB), caused by Mycobacterium tuberculosis, remains the leading cause of death

from a single infectious agent worldwide, with a disease burden surpassing that of COVID-19[1]. Although the global TB incidence has declined by 8.3% since 2015, this remains far from the World Health Organization's target of a 50% reduction by 2025[2]. Importantly, the distribution of TB is highly uneven, with approximately 95% of cases occurring in low- and middle-income countries, making these regions the primary high-burden settings for TB control[3]. Active case finding and timely, appropriate treatment have been shown to reduce incidence in such settings; however, LMICs continue to face major challenges, including poverty, overcrowding, limited health care resources, diagnostic delays, and poor treatment adherence[2,4–6].

Large Language Models (LLMs), as natural language processing tools based on large-scale pre-training and deep learning technologies, have demonstrated potential in the auxiliary diagnosis and management of various diseases[7]. Studies have shown that LLMs have potential applications in medical information retrieval, diagnostic reasoning, and treatment recommendation generation, with favorable performance reported in diverse clinical contexts such as joint pain, glaucoma, and diabetes，which can serve as useful tools to assist healthcare providers in enhancing the quality of care.[8–13]However, most existing research has focused on noncommunicable diseases such as joint pain, glaucoma, and diabetes, while evidence regarding LLM use in tuberculosis remains limited..

Early recognition, differential diagnosis, and standardized treatment of tuberculosis require efficient and reliable decision-support tools[14–16]. However, most existing TB-related artificial intelligence studies have focused on imaging, such as deep learning–based chest radio graph screening with reported sensitivities up to 95%. These tools cannot integrate multi-modal clinical information (eg, symptoms, history, laboratory tests), limiting their utility in triage, differential diagnosis, and management[17]. A 2024 systematic review noted that conversational AI may have potential for clinical decision support in low-resource settings but lacks validation in complex decision-making and raises concerns regarding algorithmic bias and data privacy[18]. These studies suggest potential value of LLMs in infectious disease management, but direct evidence in TB care is

lacking.

This study aims to evaluate the performance of large language models in tuberculosis triage, differential diagnosis, and management tasks, comparing them with tuberculosis specialists. Structured clinical cases were evaluated using consensus-based reference standards, with assessment dimensions encompassing objective metrics (precision, recall, F1 score, readability) and subjective criteria (applicability, subjective information quality, understandability, and operability). By identifying strengths and limitations, this study seeks to clarify the potential role of LLMs in tuberculosis decision support and provide evidence for their application in resource-constrained settings.

Methods

This diagnostic accuracy study was conducted to evaluate the potential of LLMs as clinical decision-support tools for tuberculosis. The study was designed and reported in accordance with the Standards for Reporting of Diagnostic Accuracy Studies (STARD). The protocol was approved for scientific validity by the Academic Committee of Clinical Research at Shenzhen Third People's Hospital (approval No. AF-MS-006-02). The study workflow is shown in eFigure 1 in Supplement 1.

Case Development and Selection

A total of 17 tuberculosis cases were included, comprising 7 virtual cases and 10 real cases. The virtual cases were developed by two tuberculosis specialists based on the 10th edition of the medical textbook *Diagnostics*, the official *WS 196–2017 Classification of Tuberculosis* issued by the Commission of the People's Republic of China, and their clinical experience, ensuring coverage of the most prevalent types of tuberculosis.[19,20]. To supplement with uncommon or complex presentations, 10 real cases were selected, representing diverse sex, age groups, and complication. All cases were presented in a structured text format, including demographic information, chief complaints, history of present illness, imaging, laboratory and microbiological findings, treatment

history, and complication (real cases only). Each case contained 2 triage questions, 1 diagnostic question, and 1 management question. Patient characteristics are summarized in Table 1, with additional details provided in eTable 1 in Supplement 1.

Table 1 . Characteristics of the Case Vignettes

	Fictional Case (N=7)	Real Case (N=10)	Participants, No. (%) Overall (N=17)
Age, median (IQR)	43 (25-49)	39 (37-61)	41 (35-51)
Gender			
Female	3 (43)	5 (50)	8 (47)
Male	4 (57)	5 (50)	9 (53)
Tuberculosis type			
Pulmonary tuberculosis	6 (83)	7 (70)	13 (76)
Extrapulmonary tuberculosis	1 (17)	3 (30)	4 (24)

Standardized Interpretation of Tuberculosis Clinical Information

The reference standard was established using an expert consensus approach by 3 senior tuberculosis specialists (chief or associate chief physicians, each with ≥10 years of clinical experience). The panel independently reviewed each structured case record and addressed 4 predefined domains: hospital type (recommending the appropriate hospital level, such as community, general, or tertiary hospitals) and triage department (recommending the appropriate clinical department based on the patient's condition), differential diagnosis, and management recommendations. Through iterative discussion, consensus was reached on all items. The finalized responses were designated as the gold standard for subsequent comparisons with outputs from LLMs and physicians.

Data collection

Before study initiation, participants were divided into 2 groups: LLMs and physicians. ChatGPT-4o and DeepSeek-R1 were selected as representative models based on prior evidence of reliable performance in medical reasoning and clinical tasks[21–25]. The physician group comprised 2 randomly selected tuberculosis specialists (attending physicians, each with ≥3 years of experience).

All participants used identical prompts requiring responses on hospital triage level and type, differential diagnosis, and treatment management (Supplement 1, eTable 1). Physicians were blinded

to LLM outputs to ensure independence between groups. To minimize temporal bias, all LLM queries were conducted by a single investigator on April 21, 2025. In this study, LLMs were defined as acting in the role of tuberculosis specialists, with outputs restricted to standard medical abbreviations without further explanations, to mimic physician communication style.

Expert rating

Given the lack of standardized tools to evaluate medical advice generated by large language models or physicians, an online information quality assessment form was developed with reference to established performance metrics for generative AI, with items selected according to the core requirements of each task[26–30]. After anonymization, 3 tuberculosis specialists evaluated 68 responses generated by 2 physicians and 2 LLMs, yielding 816 ratings (eTable 2 in Supplement 1).

The assessment included objective performance and subjective experience. Objective performance measured the accuracy and completeness of recommendations provided by LLMs or physicians across triage, diagnosis, and management tasks, quantified using precision, recall, and F1 score. Recall, also called the true positive rate, is a machine learning evaluation metric that measures the proportion of correct information that is identified out of all correct information. F1 is a machine learning evaluation metric that captures the balance between accuracy and completeness. Subjective perception measured experts' perceptions of the recommendations, including suitability and subjective information quality across all tasks, and conciseness, safety, operability, and comprehensibility for management tasks, reflecting the acceptability of the information and the potential of LLMs to assist clinical decision-making. Text readability was analyzed using Jieba 0.42 and a validated Chinese readability formula, with medical terminology processed according to the Common Clinical Medical Terms (2023 edition)[31,32].Definitions of all indicators are provided in Table 1, and detailed scoring criteria are presented in eTable2 in Supplement 1.

Table2 Indicators for Evaluating Performance and Information Quality in Triage, Differential Diagnosis, and Management Recommendation

Evaluation Module	Indicator	Definition	Method	Description
Objective Performance (accuracy and completeness of recommendations in triage, diagnosis, and management)	Accuracy	The proportion of correct answers identified by the doctor/LLM relative to all correct answers.	$Precision = \frac{TP}{TP + FP}$ TP (True Positive): Items proposed by the LLM or physician that exactly match the expert-standard answers. FP (False Positive): Items proposed by the LLM or physician that are not included in the expert-standard answers, including extra or incorrect items (misdiagnoses or errors).	Reflects the accuracy of LLM/physician judgments; higher values indicate fewer misdiagnoses or incorrect items, and more reliable judgments or recommendations.
	Completeness	The proportion of correct information that doctors/LLMs can identify out of all correct information.	$Recall = \frac{TP}{TP + FN}$ TP (True Positive): Items proposed by the LLM or physician that exactly match the expert-standard answers. FN (False Negative): Items present in the expert-standard answers but not proposed by the LLM or physician (missed items).	Reflects the ability to capture all relevant items; higher values indicate fewer missed items, more comprehensive coverage, and more complete judgments or recommendations.
	F1 score	The harmonic mean of precision and recall， used to evaluate a model's performance in terms of both accuracy and completeness.	$F1\ score = \frac{precision \times recall \times 2}{precision + recall}$	Balances accuracy and coverage; higher values indicate stronger overall performance in judgments and recommendations, making the LLM/physician more suitable as a decision-support tool.
Subjective perception (experts' perceived acceptability of information)	Suitability	Provide patients with appropriate information to assist them in correctly understanding their condition or obtaining suitable medical care, without causing ambiguity or misdirection.	Five-point Likert scale	Measures how well the information helps patients understand their condition or obtain care; higher scores indicate more appropriate and reliable guidance.
	Subjective information quality	Subjective perception of the overall quality of the text		Measures perceived overall quality; higher scores indicate users consider the information higher quality.
	conciseness	The necessity of the information		Measures necessity and brevity; higher scores indicate concise information without unnecessary content.
	Safety	The extent to which the information contains errors or hazardous information		Measures error-free and hazard-free content; higher scores indicate safer information.
	understandability	The clarity and comprehensibility of information for non-specialists		Measures clarity for non-specialists; higher scores indicate easier understanding.
	operability	The feasibility of implementing information in actual healthcare settings		Measures feasibility of applying the information in practice; higher scores indicate more practical and actionable information.
Text readability	R value	Excluding the impact of medical terminology, the difficulty of reading information provided by doctors or LLM	$R = 17.5255 \times Total\ word\ count + 0.4415 \times average\ sentence\ length - 18.3344 \times (1 - Proportion\ of\ Medical\ Terminology)$	Measures the readability difficulty of information excluding medical terminology; higher scores indicate the text is easier to read and understand.

Statistical analysis

Statistical analyses were conducted using SPSS(27.0), Python(3.13), and OriginPro(2025). The mean of 3 expert ratings was used for each response to ensure independence. Owing to non-normal data, results are reported as median (IQR). Group differences were tested with the Mann-Whitney U test (2-sided, $P < .05$), and effect size r was calculated as Z divided by the square root of the total sample size (N), with 95% *CI* obtained by bootstrap. Inter-rater reliability was evaluated using the intraclass correlation coefficient (ICC) from a 2-way random-effects model.

Results

Objective Performance Evaluation

In the triage and diagnosis tasks, F1 scores did not differ significantly between groups. In the triage task, the median (IQR) F1 score was 0.67 (0.33–0.67) for the LLM group and 0.50 (0.33–0.67)

for the physician group (r = 0.06; 95% *CI*, 0.00–0.20;*U* = 2169.0; *P* = .503). In the diagnosis task, both groups had 0.33 (0.33–0.67) (r = 0.12; 95% *CI*, 0.00–0.35; *U* = 505.0; *P* = .321) (Table 2).

In the treatment management task, the LLM group showed higher recall and F1 score. The median (IQR) recall was 0.50 (0.39–0.58) in the LLM group and 0.20 (0.12–0.31) in the physician group (r = 0.71; 95% *CI*, 0.56–0.79; *U* = 104.0; *P* < .001). The F1 score was 0.50 (0.44–0.62) vs 0.30 (0.20–0.43) (r = 0.58; 95% *CI*, 0.40–0.72; *U* = 104.0; *P* < .001). Precision was 0.60 (0.50–0.69) in the LLM group and 0.55 (0.50–0.85) in the physician group, with no significant difference (r = 0.04; 95% *CI*, 0.00–0.15; *U* = 607.0; *P* = .717) (Table 2).

For overall performance across tasks, precision did not differ. The median (IQR) precision was 0.67 (0.33–0.67) in the LLM group and 0.50 (0.33–0.67) in the physician group (*U* = 8695.5; r = 0.06; 95% *CI*, 0.00–0.15; *P* = .346). Recall was higher in the LLM group, 0.53 (0.33–0.67), compared with 0.33 (0.31–0.67) in the physician group (*U* = 6848.5; r = 0.23; 95% *CI*, 0.11–0.35; *P* < .001). F1 score was also higher in the LLM group, 0.58 (0.33–0.67), vs 0.33 (0.33–0.67) in the physician group (*U* = 7085.5; r = 0.21; 95% *CI*, 0.08–0.31; *P* < .001). These results indicate superior overall performance of the LLM, especially in recall and F1 score (Table 2).

Table3 Comparison of LLMs and Physicians for Precision, Recall, and F1-Score Across different tasks and overall

	Group		*U*	*P* value
	LLMs	Physicians		
Triage				
Precision	0.67（0.33-0.67）	0.50（0.33-0.67）	2169.0	*P* =0.503
Recall	0.67（0.33-0.67）	0.50（0.33-0.67）	2169.0	*P* =0.503
F1 score	0.67（0.33-0.67）	0.50（0.33-0.67）	2169.0	*P* =0.503
Differential diagnoses				
Precision	0.33（0.33-0.67）	0.33（0.33-0.67）	505.0	*P* =0.321
Recall	0.33（0.33-0.67）	0.33（0.33-0.67）	505.0	*P* =0.321
F1 score	0.33（0.33-0.67）	0.33（0.33-0.67）	505.0	*P* =0.321
Management recommendations				
Precision	0.60（0.50-0.69）	0.55（0.50-0.85）	607.0	*P* =0.717
Recall	0.50（0.39-0.58）	0.20（0.12-0.31）	104.0	*P*<0.001
F1 score	0.50（0.44-0.62）	0.30（0.20-0.43）	102.0	*P*<0.001
Overall Precision	0.67（0.33-0.67）	0.50（0.33-0.67）	8695.5	*P* =0.346
Overall Recall	0.53（0.33-0.67）	0.33（0.31-0.67）	6848.5	*P*<0.001
Overall F1 score	0.58（0.33-0.67）	0.33（0.33-0.67）	7085.5	*P*<0.001

Based on 17 independent cases, precision, recall, and F1-score were calculated for each case. Results are presented as median (inter quartile range), with between-group comparisons performed using the Mann-Whitney *U* test (*U* statistic and exact P-values shown; significance defined as $P < 0.05$).

Subjective preception evaluation

Three experts evaluated 68 answers based on applicability and subjective information quality (all tasks) as well as conciseness, safety, comprehensibility, and operability (management recommendations only). Interrater reliability was moderate (*ICC* = 0.54; 95% *CI*, 0.48–0.59; *P* < .001). The LLM group consistently outperformed the physician group in suitability and subjective information quality across both overall and task-specific evaluations, with statistically significant differences.

For overall performance, the median (IQR) suitability score was 3.67 (3.33–4.00) in the LLM group vs 3.00 (2.67–3.33) in the physician group (r = 0.77; 95% *CI*, 0.73–0.81; *U* = 1122.0; *P* < .001). The median (IQR) score for subjective information quality was 3.33 (3.33–3.67) vs 2.67 (2.33–2.67) (r = 0.87; 95% *CI*, 0.85–0.88; *U* = 155.0; *P* < .001)(Figure1).

In the triage task, the median (IQR) suitability score was 3.67 (3.33–4.00) for the LLM group vs 3.00 (2.67–3.33) for the physician group (r = 0.77; 95% *CI*, 0.69–0.81; *U* = 292.5; *P* < .001).

Subjective information quality was 3.33 (3.33–3.67) vs 2.67 (2.33–2.67) (r = 0.86; 95% *CI*, 0.83–0.88; *U* = 42.0; *P* < .001). In the differential diagnosis task, suitability was 3.67 (3.33–4.00) vs 3.00 (3.00–3.33) (r = 0.71; 95% *CI*, 0.55–0.80; *U* = 111.5; *P* < .001). Subjective information quality was 3.33 (3.33–3.67) vs 2.33 (2.33–2.67) (r = 0.71; 95% *CI*, 0.55–0.80; *U* = 10.5; *P* < .001).In the management recommendation task, suitability was 3.67 (3.67–4.00) vs 2.67 (2.67–3.00) (r = 0.83; 95% *CI*, 0.76–0.86; *U* = 29.5; *P* < .001). Subjective information quality was 3.33 (3.33–3.67) vs 2.33 (2.33–2.67) (r = 0.86; 95% *CI*, 0.82–0.88; *U* = 8.0; *P* < .001)(Figure1).

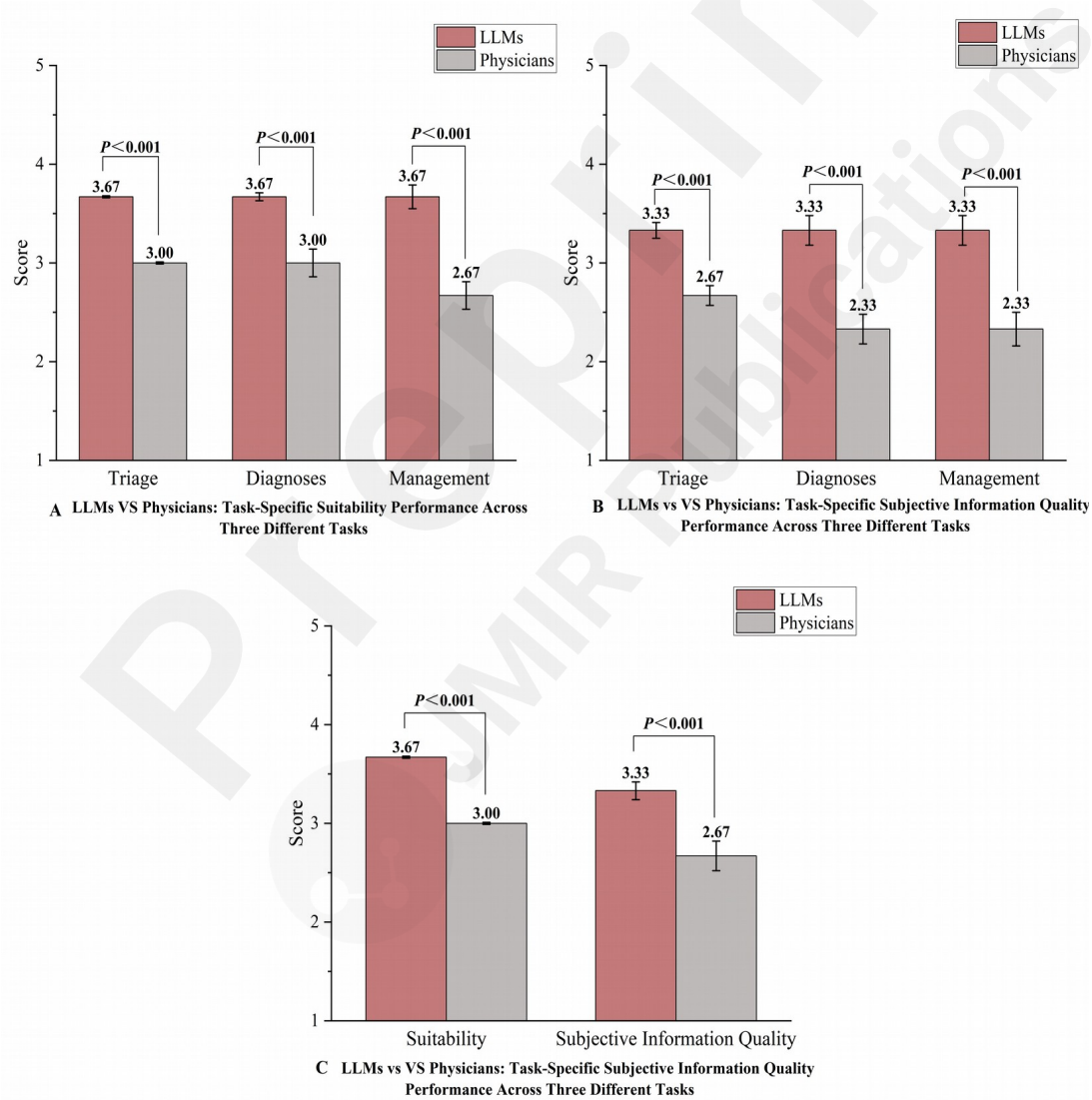


Figure1 Comparison of LLM and Physician Performance on Suitability and Subjective Information Quality Across Specific and Overall Tasks

Bars represent median values, with error bars indicating the standard error of the median (SEM) derived from bootstrap resampling (n=5,000 iterations).The significance of inter group differences was determined using the Mann-Whitney U test (significance level α=0.05).

In the tuberculosis management recommendation task, the LLM performed better than

physicians in operability and understandability, whereas no significant differences were observed in safety or conciseness.

For conciseness, the median (IQR) score was 4.00 (3.67–4.00) in the LLM group and 4.00 (3.00–4.00) in the physician group (r = 0.33; 95% *CI*, 0.10–0.52; *U* = 4965.0; *P* = .538). For safety, the median (IQR) score was 4.00 (4.00–4.30) in the LLM group and 4.00 (4.00–5.00) in the physician group (r = 0.21; 95% *CI*, 0.01–0.40; *U* = 5936.5; *P* = .061)(Figure1).

For understandability, the median (IQR) score was 3.67 (3.33–4.00) in the LLM group and 3.00 (3.00–4.00) in the physician group (r = 0.77; 95% *CI*, 0.65–0.84; *U* = 4281.5; *P* = .022). For operability, the median (IQR) score was 3.67 (3.33–4.00) in the LLM group and 3.00 (3.00–4.00) in the physician group (r = 0.79; 95% *CI*,
0.70–0.80; *U*= 4305.0; *P*= .025)(Figure2).

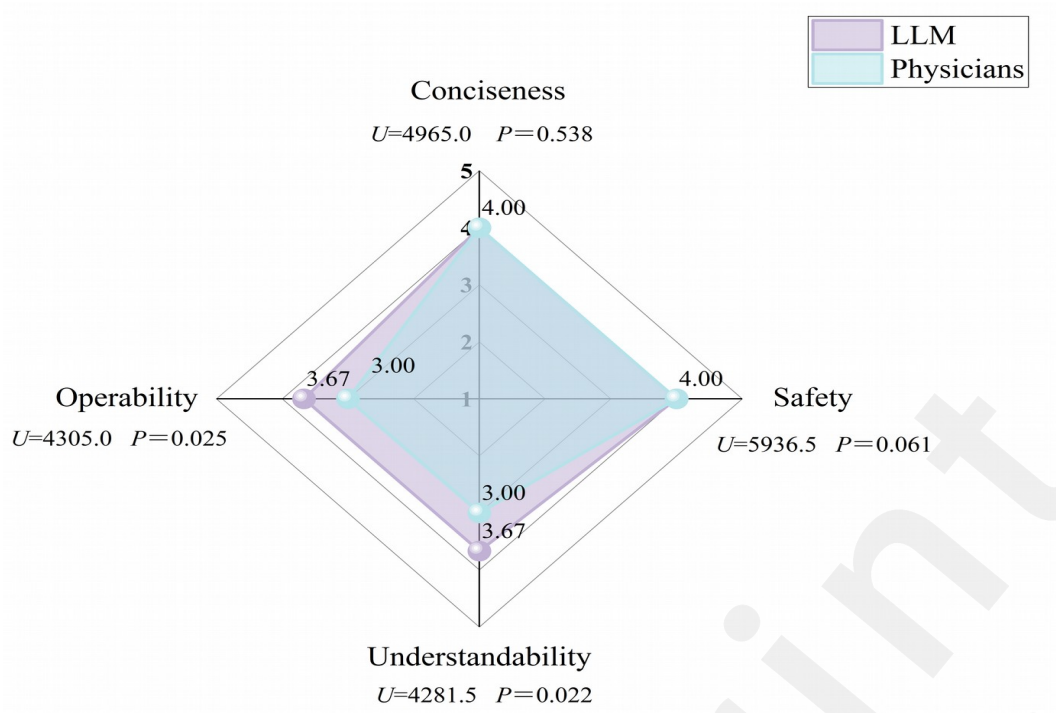


Figure2 Comparisons of LLM and Physicians: Performance in Management Recommendations Across Four Domains

This figure illustrates the median scores achieved by the LLM and physician groups across various dimensions in the management task, employing the Mann-Whitney U test to compare the differences in scores between the two groups across these dimensions.

Readability analysis

We processed 272 question-response pairs between doctors and LLMs, calculating the readability score (R value) for each doctor and LLMs response to assess textual comprehension difficulty.The median R values (interquartile range) for the LLM group and physician group texts were 2.17 (1.35–2.65) and 1.88 (0.26–2.32), respectively. A statistically significant difference existed between the two groups' readability scores (r = 0.21, 95% *CI*: 0.07–0.32, *U* = 7068.5; *P* < 0.001), indicating that LLM-generated texts present greater reading difficulty (Figure 3).

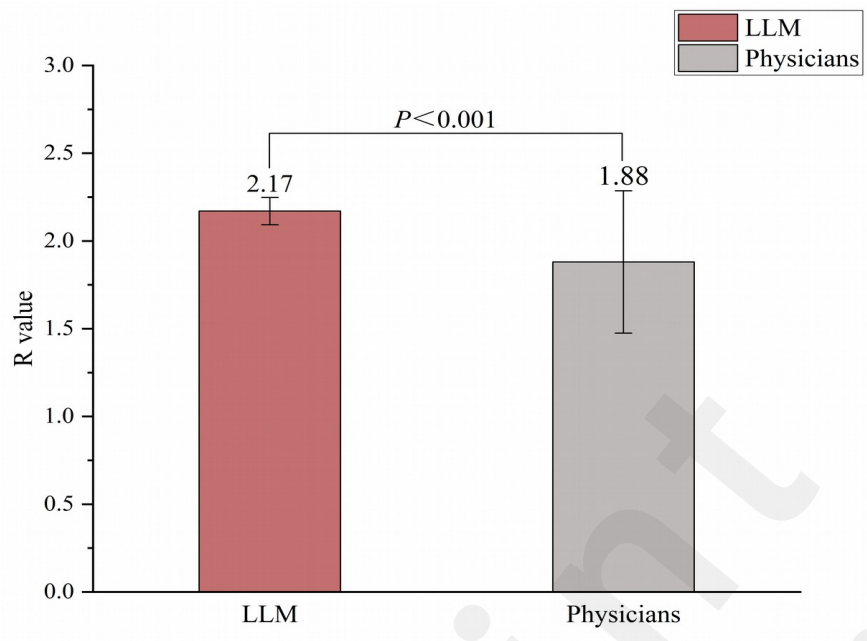


Figure3 Comparison of Readability Between LLM-Generated and Physician-Generated Content
Bars represent the median readability scores for each group. Error bars indicate the standard error of the median (SEM) derived from bootstrap resampling (n=5,000 iterations).The significance of inter group differences was determined using the Mann-Whitney U test (significance level α=0.05).

Discussion

This study investigated the potential of large language models (LLMs) as decision-support tools for tuberculosis by comparing their performance against clinicians across three core clinical tasks: triage, differential diagnosis, and management recommendations. The LLM group consistently outperformed clinicians in recall and F1 scores, especially for treatment recommendations. Higher recall indicates that LLMs were less likely to miss relevant diagnostic or management elements, while higher F1 scores reflect a stronger balance between accuracy and comprehensive coverage. LLM outputs were also rated superior in suitability, understandability, and operability, meaning their recommendations were more appropriate for care, clearer to interpret, and more feasible to apply. Precision did not differ significantly, suggesting LLMs maintain accuracy while enhancing completeness and clinical utility. Collectively, these results support the potential for LLMs to serve as clinical decision-support tools for tuberculosis.

These findings are consistent with previous literature, indicating that LLMs excel at handling complex medical tasks, particularly in information retrieval and synthesis[23,28,29,33,34]. In tuberculosis

care, the advantage of higher recall is particularly critical, as missed diagnoses frequently lead to treatment delays and sustained transmission[35,36]. LLMs can generate standardised, structured management recommendations following diagnosis, helping to prevent critical management steps from being overlooked. This ensures continuity of care and standardised management for patients. In primary healthcare settings, this function is particularly crucial, as non-specialist practitioners often undertake the primary diagnostic and therapeutic responsibilities. In such contexts, structured recommendations serve as a safety net, reminding less experienced clinicians not to overlook critical information. Moreover, large language models demonstrate comparable accuracy to medical practitioners, indicating they do not compromise precision and can serve as a reliable supplementary information source. In resource-constrained regions, these models hold promise in alleviating the pressure of specialist shortages, providing high-quality decision support for primary healthcare workers.

Concurrently, this study suggests that the outstanding independent performance of LLMs does not imply that physicians can directly utilise their outputs to enhance reasoning capabilities. Oncology research indicates that even with LLM suggestions, physicians' diagnostic accuracy may not improve[37]. One reason lies in LLMs' high sensitivity to prompt words; without structured input frameworks, clinicians struggle to consistently leverage their advantages[38,39]. Consequently, for tuberculosis clinical applications, standardised prompts and built-in query templates must be developed alongside enhanced physician training to ensure effective implementation.

Regarding readability, our findings diverge from previous studies[32]. By incorporating medical terminology difficulty into our assessment, we observed that physician-generated texts proved more comprehensible to non-specialists. This suggests that while LLMs demonstrate clinical logical accuracy, improvements remain necessary for patient communication and health education. Future enhancements could focus on terminology simplification, explanatory annotations, or adaptive

language strategies to boost acceptability—maintaining professionalism while enhancing intelligibility.

In terms of study design, this research employed a restricted output framework for triage and differential diagnosis tasks, limiting LLMs to a maximum of three candidate answers in diagnostic scenarios. Consequently, precision could only assume four discrete values (0, 0.33, 0.67, and 1), thereby reducing variability. While this design helped focus on the LLM's foundational reasoning capabilities, it also limited differentiation between over diagnosis and under diagnosis. Similar phased designs have been applied in LLM research within orthopaedics[10], serving as crucial exploratry steps before advancing to complex real-world scenarios.

In summary, this study demonstrates that LLMs outperform clinicians in triage, differential diagnosis, and management recommendations, exhibiting particularly significant advantages in recall, F1 score, and subjective usability while maintaining comparable accuracy to physicians. This suggests their potential as clinical decision support tools for tuberculosis. However, achieving clinical translation requires optimising readability, enhancing clinician training, and promoting deep integration with clinical workflows. Concurrently, key challenges including liability attribution, privacy protection, and quality control remain pressing issues requiring resolution.

Limitations

This study has several limitations. First, the restricted output framework for triage and diagnostic tasks resulted in identical false positive and false negative values, limiting differentiation between overdiagnosis and missed diagnosis. In tuberculosis, this distinction is clinically important because overdiagnosis increases healthcare burden, while missed diagnosis delays treatment and fosters transmission. Second, the relatively small sample size constrains generalizability, underscoring the need for larger, multi-center studies. Third, although readability assessment considered terminology difficulty, it was not aligned with conventional indices such as the Flesch-

Kincaid grade, limiting comparability. Finally, real-world integration of LLMs was not evaluated; challenges including accountability, interoperability, and data security remain critical. Future studies should address these issues to clarify the role of LLMs in tuberculosis care.

Conclusions

This study demonstrates that large language models (LLMs) show strong potential as auxiliary tools in tuberculosis care. LLMs matched physicians in accuracy while achieving higher recall and F1 scores, meaning they captured more relevant diagnostic and management elements with balanced correctness. These strengths reduce the risk of overlooking critical clinical information and enable more comprehensive decision-making, indicating that LLMs can enhance early recognition and treatment planning without sacrificing precision. Subjective evaluations indicate LLMs possess superior advantages in information delivery quality. However, the readability of model-generated text falls below that of clinicians' output, indicating further optimisation is required for practical implementation. These findings confirm that large language models hold promise as critical adjuncts to clinical decision-making in tuberculosis, enhancing diagnostic efficiency and decision quality. Future research should focus on expanding sample sizes, conducting multi-centre validation, and developing terminology interpretation capabilities. This will improve the comprehensibility of LLMs in real-world clinical settings, thereby fully unlocking their value in supporting tuberculosis diagnosis and treatment decision.

Acknowledgments

We would like to express our sincere gratitude to the research team members who participated in data collection, evaluation, and statistical analysis.

This study was designed and conducted by Jin Liao, Wenjun He, and Jinghui Chang. Lanping Zhang and Jiamin Huang contributed to the development of the cases; Lanping Zhang additionally assisted in data organization, quality control, and preliminary analysis. Jiamin Huang and Xue Li participated in answering the cases. Zhichao Liu, Xue Ke, and Jian Li were involved in the screening

and scoring of the cases. Huiyi Pan, Jin Liao, Wenjun He, Xingyan Li, and Haiting Cai collected and organized the data, conducted quality control, and assisted in preliminary data analysis. Jin Liao conducted data screening and analysis, and wrote the first draft of the manuscript. Candice Hwang revised the manuscript. Jinghui Chang and Guobao Li contributed to the supervision of data analysis and manuscript preparation. Jin Liao and Wenjun He are joint first authors, and Jinghui Chang and Guobao Li are corresponding authors. All authors read and approved the final manuscript.

Funding

This work was supported by Southern Medical University Center for World Health Organization Studies, Research Base for Development of Public Health Service System of Guangzhou, Key Laboratory of Philosophy and Social Sciences of Guangdong Higher Education Institutions for Health Polices Research and Evaluation [grant number 2015WSY0010], the Youth Project of Guangdong Provincial Philosophy and Social Sciences Planning Project [grant number GD25CGL23], Guangdong Humanities and Social Sciences Popularization Base for Public Health Emergency and Health Education, and the General Project of Guangdong Provincial Medical Research Fund [grant number A2025059].

Conflicts of Interest

No conflict.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Abbreviations

Large Language Model LLM

Reference

1. *World Health Organization. Tuberculosis resurges as top infectious disease killer. Published October 29, 2024. Accessed April 13, 2025 . https://www.who.int/zh/news/item/29-10-2024-tuberculosis-resurges-as-top-infectious-disease-killer*

2. *Global Tuberculosis Report 2024*. 1st ed. World Health Organization; 2024.

3. Cioboata R, Balteanu MA, Osman A, et al. Coinfections in tuberculosis in low- and middle-income countries: epidemiology, clinical implications, diagnostic challenges, and management strategies—a narrative review. *J Clin Med*. 2025;14(7):2154. doi:10.3390/jcm14072154

4. Költringer FA, Annerstedt KS, Boccia D, Carter DJ, Rudgard WE. The social determinants of national tuberculosis incidence rates in 116 countries: a longitudinal ecological study between 2005-2015. *BMC Public Health*. 2023;23(1):337. doi:10.1186/s12889-023-15213-w

5. Xie LL. From inequity to equity: progress of research on medical equity among rural residents. Jianghan Academic. 2021;40(5):83-92. doi:10.16388/j.cnki.cn42-1843/c.2021.05.008

6. Yuen CM, Amanullah F, Dharmadhikari A, et al. Turning off the tap: stopping tuberculosis transmission through active case-finding and prompt effective treatment. *Lancet Lond Engl*. 2015;386(10010):2334-2343. doi:10.1016/S0140-6736(15)00322-0

7. Hao J, Yao Z, Tang Y, Remis A, Wu K, Yu X. Artificial intelligence in physical therapy: evaluating ChatGPT's role in clinical decision support for musculoskeletal care. *Ann Biomed Eng*. 2025;53(1):9-13. doi:10.1007/s10439-025-03676-4

8. Cascella M, Montomoli J, Bellini V, Bignami E. Evaluating the feasibility of ChatGPT in healthcare: an analysis of multiple clinical and research scenarios. *J Med Syst*. 2023;47(1):33. doi:10.1007/s10916-023-01925-4

9. Savage T, Nayak A, Gallo R, Rangan E, Chen JH. Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine. *Npj Digit Med*. 2024;7:20. doi:10.1038/s41746-024-01010-1

10. Kunze KN, Varady NH, Mazzucco M, et al. The Large Language Model ChatGPT-4 Exhibits Excellent Triage Capabilities and Diagnostic Performance for Patients Presenting With Various Causes of Knee Pain. *Arthrosc J Arthrosc Relat Surg Off Publ Arthrosc Assoc N Am Int Arthrosc Assoc*. Published online June 24, 2024:S0749-8063(24)00456-0. doi:10.1016/j.arthro.2024.06.021

11. Laghari M, Talpur BA, Sulaiman SAS, Khan AH, Bhatti Z. Assessment of adherence to anti-tuberculosis treatment and predictors for non-adherence among the caregivers of children with tuberculosis. *Trans R Soc Trop Med Hyg*. 2021;115(8):904-913. doi:10.1093/trstmh/traa161

12. Qu RW, Qureshi U, Petersen G, Lee SC. Diagnostic and management applications of ChatGPT in structured otolaryngology clinical scenarios. *OTO Open*. 2023;7(3):e67. doi:10.1002/oto2.67

13. Liu J, Wang C, Liu S. Utility of ChatGPT in clinical practice. *J Med Internet Res*. 2023;25:e48568. doi:10.2196/48568

14. Tuberculosis (TB). Accessed April 14, 2025. https://www.who.int/news-room/fact-sheets/detail/tuberculosis

15. Datta S, Evans CA. The uncertainty of tuberculosis diagnosis. *Lancet Infect Dis*. 2020;20(9):1002-1004. doi:10.1016/S1473-3099(20)30400-X

16. Munro SA, Lewin SA, Smith HJ, Engel ME, Fretheim A, Volmink J. Patient adherence to

tuberculosis treatment: a systematic review of qualitative research. *PLoS Med*. 2007;4(7):e238. doi:10.1371/journal.pmed.0040238

17. Khan FA, Majidulla A, Tavaziva G, et al. Chest x-ray analysis with deep learning-based software as a triage test for pulmonary tuberculosis: a prospective study of diagnostic accuracy for culture-confirmed disease. *Lancet Digit Health*. 2020;2(11):e573-e581. doi:10.1016/S2589-7500(20)30221-1

18. Wu J, Ma Y, Wang J, Xiao M. The application of ChatGPT in medicine: a scoping review and bibliometric analysis. *J Multidiscip Healthc*. 2024;17:1681-1692. doi:10.2147/JMDH.S463128

19. National Health Commission of the People's Republic of China. WS 196—2017 Classification of tuberculosis. Chinese Journal of Infection Control. 2018;17(4):367-368. doi:10.3969/j.issn.1671-9638.2018.04.019

20. Wan XH, Lu XF. Diagnostics. 10th ed. Beijing, China: People's Medical Publishing House; 2025. Accessed August 23, 2025.

21. Wu X, Cai G, Guo B, et al. A multi-dimensional performance evaluation of large language models in dental implantology: comparison of ChatGPT, DeepSeek, grok, gemini and qwen across diverse clinical scenarios. *BMC Oral Health*. 2025;25(1):1272. doi:10.1186/s12903-025-06619-6

22. Zhao W, Lai H, Pan B, et al. Assessing the adherence of large language models to clinical practice guidelines in chinese medicine: a content analysis. *Front Pharmacol*. 2025;16:1649041. doi:10.3389/fphar.2025.1649041

23. Tordjman M, Liu Z, Yuce M, et al. Comparative benchmarking of the DeepSeek large language model on medical tasks and clinical reasoning. *Nat Med*. 2025;31(8):2550-2555. doi:10.1038/s41591-025-03726-3

24. Chan L, Xu X, Lv K. DeepSeek-R1 and GPT-4 are comparable in a complex diagnostic challenge: a historical control study. *Int J Surg Lond Engl*. 2025;111(6):4056-4059. doi:10.1097/JS9.0000000000002386

25. Patel A, Ruoff C, Helgeson SA, Carvalho DZ, Castillo PR, Cheung J. Diagnostic performance of large language models (LLMs) compared with physicians in sleep medicine. *Sleep Med*. 2025;134:106677. doi:10.1016/j.sleep.2025.106677

26. Benary M, Wang XD, Schmidt M, et al. Leveraging Large Language Models for Decision Support in Personalized Oncology. *JAMA Netw Open*. 2023;6(11):e2343689. doi:10.1001/jamanetworkopen.2023.43689

27. Yalamanchili A, Sengupta B, Song J, et al. Quality of Large Language Model Responses to Radiation Oncology Patient Care Questions. *JAMA Netw Open*. 2024;7(4):e244630. doi:10.1001/jamanetworkopen.2024.4630

28. Frosolini A, Catarzi L, Benedetti S, et al. The Role of Large Language Models (LLMs) in Providing Triage for Maxillofacial Trauma Cases: A Preliminary Study. *Diagnostics*. 2024;14(8):839. doi:10.3390/diagnostics14080839

29. Yau JYS, Saadat S, Hsu E, et al. Accuracy of prospective assessments of 4 large language model chatbot responses to patient questions about emergency care: experimental comparative study. *J Med Internet Res*. 2024;26:e60291. doi:10.2196/60291

30. Kunze KN, Varady NH, Mazzucco M, et al. The large language model ChatGPT-4 exhibits excellent triage capabilities and diagnostic performance for patients presenting with various causes of knee pain. *Arthrosc J Arthrosc Relat Surg Off Publ Arthrosc Assoc N Am Int Arthrosc Assoc*. 2025;41(5):1438-1447.e14. doi:10.1016/j.arthro.2024.06.021

31. *National Health Commission of the People's Republic of China. Notice on the publication of commonly used clinical medical terms (2023 edition). Published March 2023. Accessed August 11, 2025.* *https://www.nhc.gov.cn/yzygj/c100068/202403/83455f24ebd54015b5d14532e072a833.shtml*

32. *Qin Q, Ke Q, Ding SY. Readability calculation and empirical application of Chinese online health education information: a case study in food safety.* Modern Information. *2020;40(5):111-121. doi:10.3969/j.issn.1008-0821.2020.05.014*

33. Gaber F, Shaik M, Allega F, et al. Evaluating large language model workflows in clinical decision support for triage and referral and diagnosis. *Npj Digit Med*. 2025;8(1):263. doi:10.1038/s41746-025-01684-1

34. Goh E, Gallo R, Hom J, et al. Large language model influence on diagnostic reasoning. *JAMA Netw Open*. 2024;7(10):e2440969. doi:10.1001/jamanetworkopen.2024.40969

35. Miller AC, Arakkal AT, Koeneman S, et al. Incidence, duration and risk factors associated with delayed and missed diagnostic opportunities related to tuberculosis: a population-based longitudinal study. *BMJ Open*. 2021;11(2):e045605. doi:10.1136/bmjopen-2020-045605

36. Storla DG, Yimer S, Bjune GA. A systematic review of delay in the diagnosis and treatment of tuberculosis. *BMC Public Health*. 2008;8(1):15. doi:10.1186/1471-2458-8-15

37. Benary M, Wang XD, Schmidt M, et al. Leveraging large language models for decision support in personalized oncology. *JAMA Netw Open*. 2023;6(11):e2343689. doi:10.1001/jamanetworkopen.2023.43689

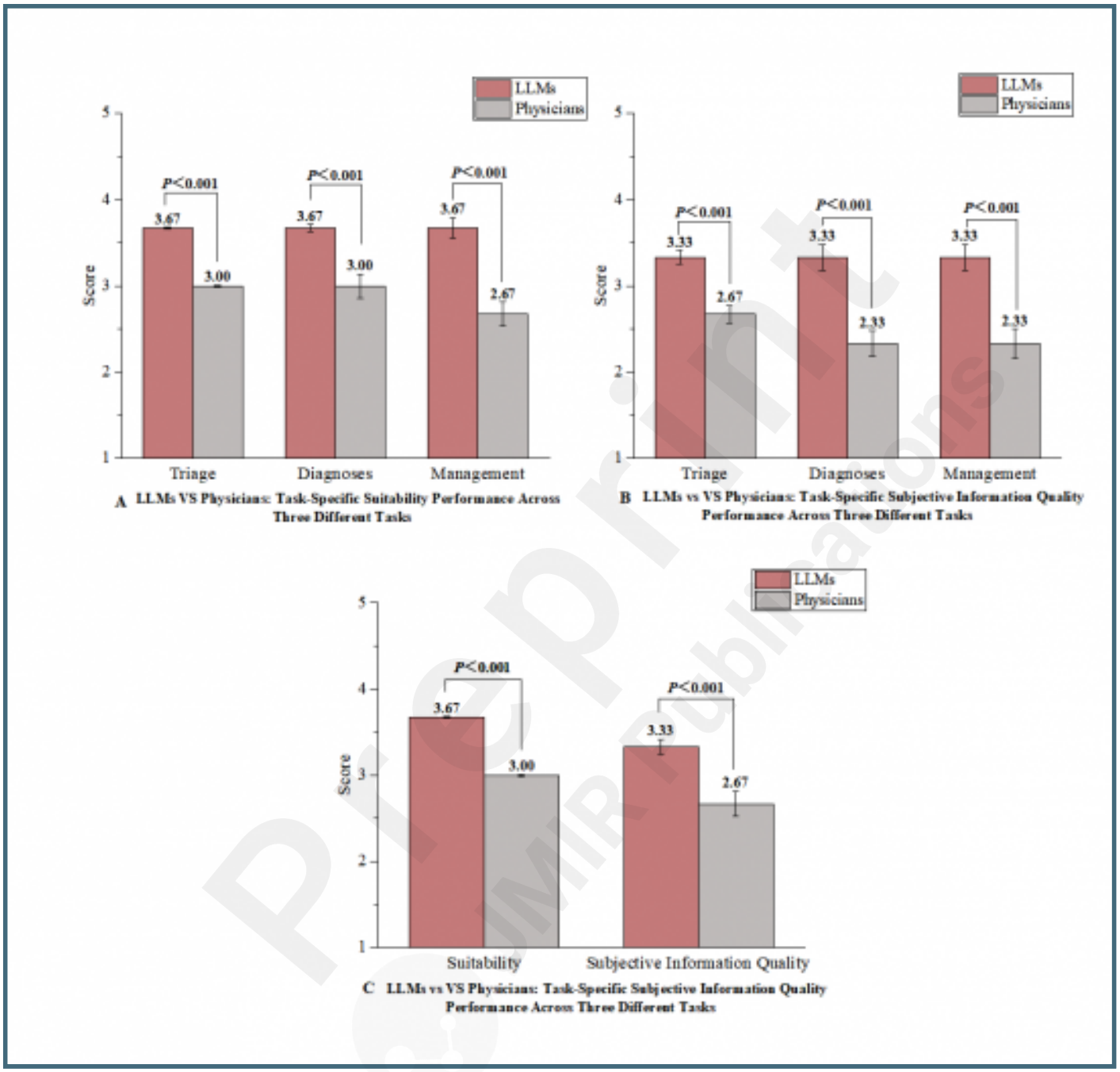
38. Arroyo AMC, Munnangi M, Sun J, et al. Open (Clinical) LLMs are Sensitive to Instruction Phrasings. *arXiv*. Preprint posted online July 12, 2024. doi:10.48550/arXiv.2407.09429

39. Sivarajkumar S, Kelley M, Samolyk-Mazzanti A, Visweswaran S, Wang Y. An empirical evaluation of prompting strategies for large language models in zero-shot clinical natural language processing: algorithm development and validation study. *JMIR Med Inform*. 2024;12(1):e55318. doi:10.2196/55318

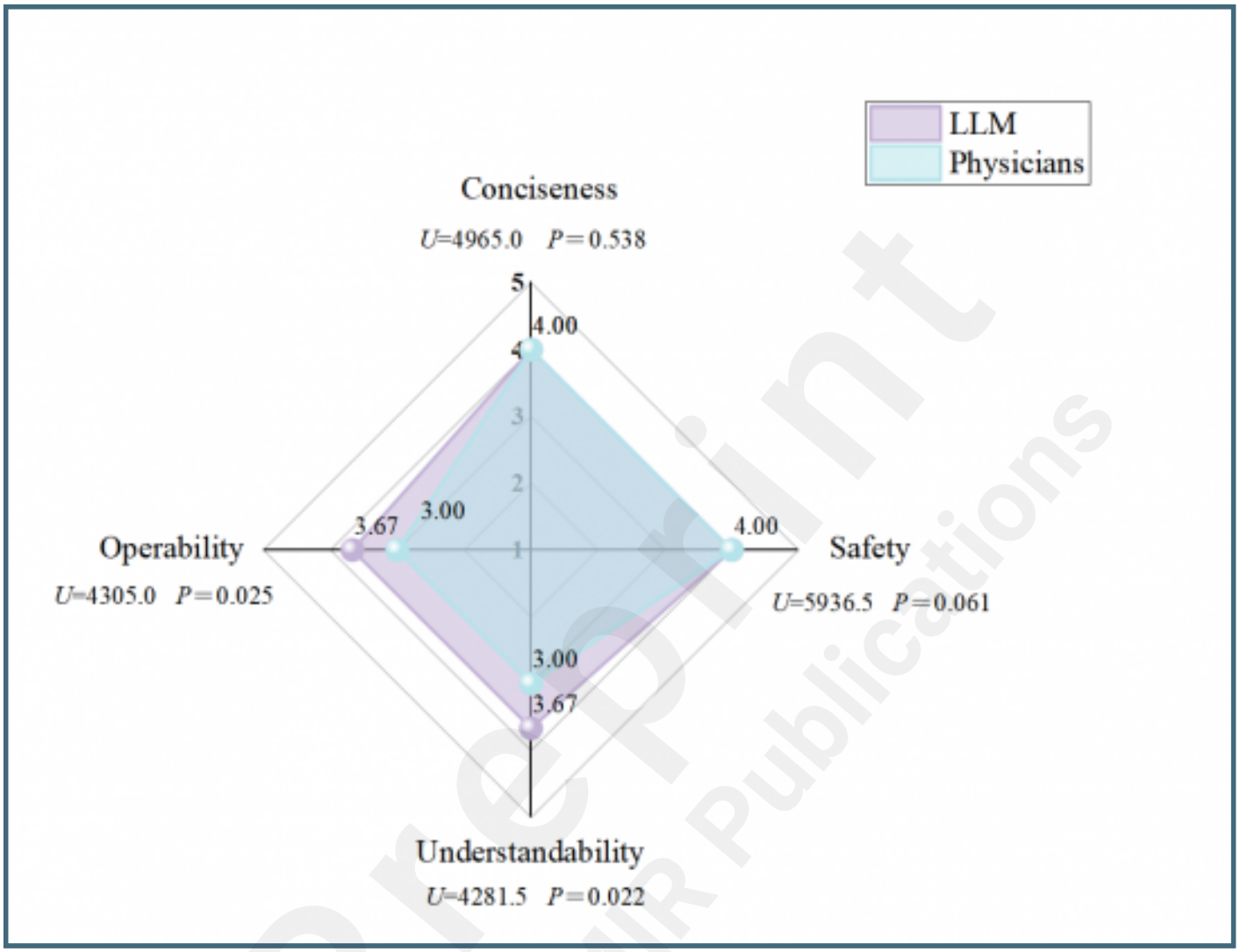
Supplementary Files

Figures

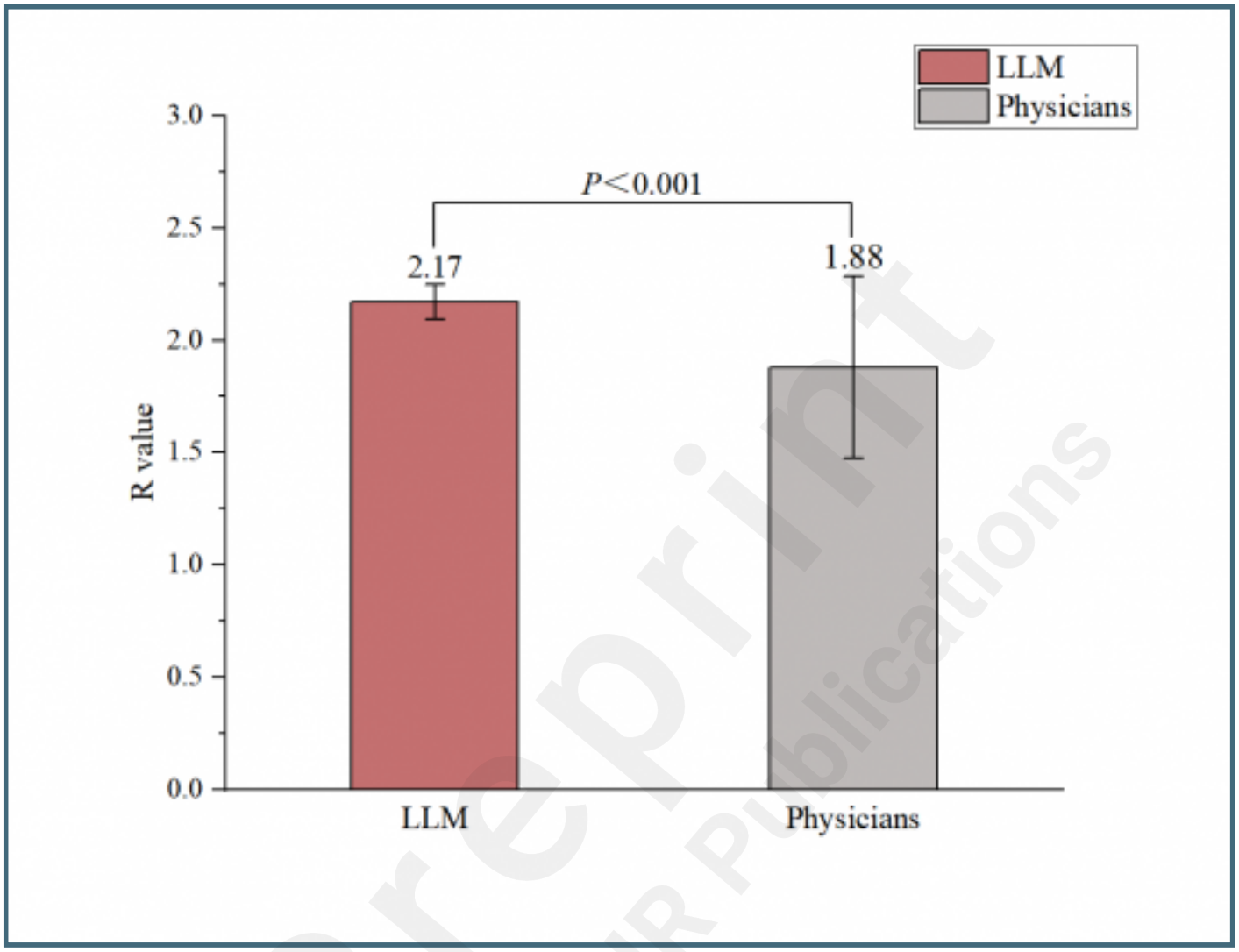
Comparison of LLM and physician performance on suitability and subjective information quality across specific and overall tasks.



Comparisons of LLM and physicians: performance in management recommendations across four domains.



Comparison of readability between LLM-generated and physician-generated content.



Multimedia Appendixes

Supplementary material: participant characteristics, LLM prompts, and scoring criteria.
URL: http://asset.jmir.pub/assets/1918d05d6ac2d697ee789d331a41f7b6.docx