Natural Resources for Human Health
eISSN: 2583-1194

Original Research

Received 25 July 2026
Revised 22 Aug 2026
Accepted 08 Sep 2026

Natural Resources for Human Health 2026, Vol. 6 (22s): 92–103

KEYWORDS:
large language models; clinical decision support; differential diagnosis; diagnostic accuracy; clinician–AI collaboration; patient safety; algorithmic bias; narrative review.

Large Language Models for Clinical Diagnostic Decision Support: A Critical Structured Narrative Review of Diagnostic Performance, Clinician–AI Collaboration, Safety, Bias, and Clinical Readiness

Awadalla Abdelwahid[1*], Fatima Habib Allah Salama[2], Dalia Omar[3], Rana Fath Elrhman[4], Abdelazim Ahmed Idris[5], Dalya I A Mohamed[6], Dina Mahmoud Fathi[7], Mayada Farah Obaid[8], Hager Ahmed Elhabo Mohamed[9], Khalid Yousif Fakhri[10]

[1]Department of obstetrics and Gynecology, Alneelain University, Khartoum, Sudan
Email: awad336@yahoo.com
ORCID ID:0009-0008-3120-2786
[2]Consultant of Obstetrics and Gynecology, Adele Medical Clinics Complex, Riyadh, Saudi Arabia.
fatima19801111@gmail.com
[3]Consultant of Obstetrics and Gynecology, Alsaedy Hospital, Al Jumom,Makkah Al mukarramah, Saudi Arabia.
Drdaliaomar18@gmail.com
[4]Department of Obstetrics and Gynecology, Dallah Namar Hospital, Riyadh, Saudi Arabia.
ranajafar41@gmail.com
[5]Obstetrics&Gynecology Specialist, Jazan Health Cluster, Jazan, Saudi Arabia.
d.abazim@gmail.com
[6]Assistant professor of Clinical Anatomy, Vision colleges in Riyadh, Saudi Arabia
dalyaibrahimahmed@gmail.com
[7]Obstetrics and Gynaecology Specialist, Tabuk Health Cluster, Tabuk, Saudi Arabia
dina.mutwalli@gmail.com
ORCID ID : 0009-0009-1361-9332
[8]Pulmonolgy Senior Registrar, Altawiin Medical Group, Alkharj
mayada_farah@yahoo.com
[9]Obstetrics and Gynaecology in AL Qassim Rigon, Riyadh Al Khabra General Hospital, Saudi Arabia.
hjr1240@gmail.com
[10]Obstetrics and Gynecology MD, Sharora General Hospital, Najran Sharora, Saudi Arabia
k.y.fakhri@gmail.com

ABSTRACT: Background: Large language models (LLMs) are promising on diagnostic tasks but benchmark success alone does not mean clinical benefit; evaluation must also consider clinician collaboration, safety and equity.

Natural Resources for Human Health
eISSN: 2583-1194

Objective: To critically interpret selected empirical evidence on diagnostic performance, clinician–AI interaction, experimental safety, algorithmic bias and requirements for supervised clinical implementation (benchmarking performance vs. patient-level benefit) separately.

Methods: We performed a structured narrative synthesis of an author-curated corpus of 20 original investigations (2023–2026) by diagnostic task, comparator, endpoint, safety stress test and applicability. A dated reproducible search history, screening log and formal risk-of-bias assessment were unavailable; systematic-review completeness and new pooled effects are not claimed.

Results: In 52 selected complex cases, GPT-4 matched the reference diagnosis at rank one in 31 (60%), top five in 42 (81%) and top ten in 43 (83%). In a randomized vignette trial with 50 physicians median diagnostic reasoning score was 76% with GPT-4 access vs. 74% with conventional resources (adjusted difference −2 percentage points; 95% CI −4 to 8; P = 0.60); exploratory standalone GPT-4 analysis scored 92%. Another experiment showed differential coverage of top ten diagnoses 51.7% AMIE assisted clinician's vs 36.1% without AMIE assistance. Under deliberately adversarial conditions hallucinations occurred in 50.0%–82.7% model/case format combinations and out of 1,504 red-team outputs only 303 (20.1%) were inappropriate. These metrics did not measure harm to patients.

Conclusions: Some LLMs can find correct diagnoses in benchmark cases and some structured clinician-assistance workflows improve task-specific differential coverage. But a randomized physician experiment showed no gain from GPT-4 access; adversarial findings are not estimates of real-world patient harm; evidence does not establish clinical safety or effectiveness of autonomous diagnosis. Prospective, clinician supervised studies should measure serious diagnostic misses, time critical triage, patient outcomes fairness and performance after model updates.

1. INTRODUCTION

Diagnostic reasoning combines symptom trajectories, examination findings, prior probabilities, investigations and new information. Diagnostic error causes harm avoidable to patients and unnecessary use of health-systems. The National Academies framed diagnostic safety as a systems problem (1); implementation studies of conventional clinical decision support have identified workflow, maintenance and integration barriers (2). LLMs can summarize patient narratives; generate differential diagnoses; propose discriminating tests; respond to follow up questions but success on medical knowledge tests or fully revealed case reports should not be confused with clinical effectiveness [3,4]. In practice time-sensitive disease may be prioritized, uncertain findings responded to, escalation when unresolved and so on more important than including the final diagnosis somewhere in a long list. Published case based investigations show how study design matters: Hirosawa et al. [5], Kanjee et al. [6] and Ríos-Hoyo et al. [7] had different collections of difficult cases and different definitions of diagnostic success. Cabral et al. [8], Hager et al. [9], Feldman et al. [10], Dinc et al. [11] used different collections of difficult cases and different definitions of [11], and Rao et al. [12] tested reasoning or differential construction when information and task demand varied. Thus, reported percentages are not equivalent to real-world diagnostic safety estimates. The distinction between standalone LLM output vs clinician augmentation is also important: a randomized comparison of GPT-4 access versus conventional resources did not find physician benefit (13), but experiments with a conversational diagnostic system (AMIE) found benefits in certain simulated interactions and differential diagnosis tasks [14,15]. The difference may be interface, clinical context, comparator resources and outcome definition rather than model–physician ranking. Safety and equity go beyond diagnostic agreement: seeded false facts, red team prompts and guideline retrieval tests reveal hallucination, omission and inappropriate response hazards [16–18]; demographic substitutions and imposed diagnostic anchors shift model outputs [19–21]; translation-controlled assessments show

language sensitivity but English–Polish results cannot be used to estimate Arabic reliability [22]. A 2025 systematic review and meta-analysis pooled diagnostic performance across 83 studies, but found substantial heterogeneity (lower than expert physicians on its pooled comparison) [23]. The present structured narrative review has a narrower explanatory goal: connect diagnostic performance to clinician interaction, stress-test safety, fairness and implementation readiness rather than produce a pooled estimate. Discussion also includes the WHO governance guidance for health-related large multimodal models [24].

2. REVIEW QUESTION AND OBJECTIVES

Review question: Across real or simulated clinical diagnostic tasks what does the selected evidence show about performance of standalone LLMs (or clinician-assisted LLMs) relative to a reference diagnosis, conventional resources, clinicians or other systems? What constraints limit inference about safety, fairness and practical readiness? Primary objective. To assess clinical diagnostic performance and how much experimental evidence exists that can support supervised diagnostic use. Secondary objectives. Distinguish first-rank accuracy from top-k coverage and sequential reasoning; summarize controlled clinician assistance findings; characterize adversarial, guideline, anchoring and triage risks; demographic and multilingual robustness; priorities for prospective evaluation.

3. REVIEW DESIGN AND REPORTING APPROACH

This structured narrative review reports on an author-curated set of 20 original studies published in 2023–2026 (explanatory critical synthesis, not effectiveness estimate). We follow the principles of SANRA [26]: evidence provenance is reported; outcome constructs are distinguished; interpretive limitations are assessed; conflicting findings are presented. We do not claim PRISMA 2020 compliance [25], prospective protocol registration, exhaustive retrieval, dual screening or formal certainty-of-evidence grading. 3. Scope and study eligibility. The scope was defined as peer-reviewed original investigations of LLMs performing clinical final diagnosis, differential diagnosis, sequential reasoning, diagnostic test or urgency decisions, clinician assisted diagnostic reasoning or directly related diagnostic safety and fairness tasks. Eligible types of evidence included de-identified clinical data, published or previously unpublished diagnostic cases, physician experiments, simulated consultations, counterfactual or red team vignette audits. Educational multiple choice examination performance without a clinical diagnostic task, reviews, commentaries and protocols were outside scope. These are the criteria used to describe the curated evidence set; a prospectively applied fully logged selection protocol is not asserted.

3.3. Literature identification and provenance of evidence set

The evidence set consists of 20 identifiable primary reports [5–22,29,30] selected to cover clinical diagnostic accuracy experiments, clinician assistance experiments, sequential reasoning, adversarial safety experiments, demographic fairness experiments and multilingual performance experiments. Original search strategies (execution dates), record exports, duplicate counts, screened records and exclusion reasons were not documented so study-selection counts and PRISMA flow diagram cannot be reconstructed properly. Set is purposive and may contain relevant positive or negative studies. In subsequent updates investigators should prospectively log databases, exact strategies, coverage dates, screening decisions, model versions and data-extraction decisions.

3.4. Evidence organization and outcome definitions

We extracted and compared the units of analysis published (patients, cases, physicians or generated responses), case source, intervention/model version, comparator, information supplied, reference standard, reported effect or accuracy metric, follow-up and major limitations (Tables 1A–1B). Rank-one diagnosis, top-k coverage, reasoning rubric scores, urgency/triage decisions, guideline fidelity and output appropriateness were treated as non-interchangeable endpoints. Outputs from the same cases were not taken to be independent patient observations. Findings were integrated narratively without recalculating primary-study estimates.

3.5. Appraisal and synthesis

Evidence was critically appraised narratively for spectrum and selection bias, benchmark contamination, reference-standard quality, prompt/model version transparency, comparator validity, outcome adjudication, missing

information and clinical applicability. No formal study-specific QUADAS-2, PROBAST+AI or GRADE ratings [27,28] were performed and therefore we make no formally graded certainty claims. Controlled clinician-assistance studies were distinguished from model-only benchmarking; adversarial error proportions were distinguished from patient-harm incidence.

3.6. Ethics

This review only analyzed previously published studies and did not recruit new participants or access identifiable individual patient records; additional research ethics approval was not required for the review itself.

4. RESULTS

4.1 Overview of the included evidence. Twenty original studies were reported [5–22,29,30], published diagnostic challenges, previously unpublished cases MIMIC records randomized or controlled clinician experiments simulated consultations repeated-output safety and fairness audits Samples vary greatly in unit of analysis clinical prevalence information available scoring standard comparator. Evidence matrices (Tables 1A and 1B) organize not statistically pool effects reported. Publication year is not equivalent to date model version evaluated.

Table 1A. Included studies: diagnostic case series, system comparisons, and clinician collaboration (studies 1–10).

Study	Study design, population, and unit of analysis	Reported task-specific outcome (as published)	Critical limitation / clinical applicability
Hirosawa [5] (2023)	52 published complex cases; GPT-4	31/52 (60%) rank 1; 42/52 (81%) top 5; 43/52 (83%) top 10.	Selected published cases; cumulative ranking thresholds.
Kanjee [6] (2023)	70 diagnostic-challenge cases; generative AI	27/70 (39%) correct first diagnosis; 45/70 (64%) included correct diagnosis in list.	Highly selected cases; ordinal quality ratings.
Ríos-Hoyo [7] (2024)	75 complex published cases; LLM evaluation	22% correct first diagnosis; 68% reference diagnosis anywhere in differential.	Publication-derived cases; broader list metric.
Cabral [8] (2024)	20 staged cases; physician comparison	Compared clinical reasoning rubric performance across AI and physicians.	Small task-specific case set; scoring construct.
Hager [9] (2024)	2,400 MIMIC-derived abdominal cases	Highlighted failures of diagnosis, laboratory interpretation, guideline compliance, and instruction following.	Four abdominal pathology groups; simulated workflow.
Goh [13] (2024)	50 physicians randomized; vignette trial	Physician score 76% LLM access vs 74% control (P=.60); exploratory GPT-4-alone median 92%.	Vignette reasoning—not downstream patient outcomes.
Zack [19] (2024)	Counterfactual demographic substitutions	Detected race- and gender-associated changes in selected healthcare outputs.	Prompt-dependent audit; not patient-level disparity estimate.
Tu [14] (2025)	159 simulated cases; 20 primary care physicians	AMIE rated superior on 30/32 specialist axes and 25/26 patient-actor axes.	Text-only consultations with trained patient-actors.

Study	Study design, population, and unit of analysis	Reported task-specific outcome (as published)	Critical limitation / clinical applicability
McDuff [15] (2025)	302 published challenging cases; 20 clinicians	Clinician top-10 coverage 51.7% AMIE support vs 36.1% unassisted; search comparison 44.4%.	Published cases; intervention and assessment constraints.
Feldman [10] (2025)	36 previously unpublished cases	Top-25 without labs: DXplain 20/36, GPT-4 15/36, Gemini 14/36; with labs: 26/36, 23/36, 21/36.	Small sample; top-25 is permissive endpoint.

Table 1B. Included studies: staged reasoning, experimental safety, fairness, language, and emerging practice evaluations (studies 11–20).

Study	Study design, population, and unit of analysis	Reported task-specific outcome (as published)	Critical limitation / clinical applicability
Dinc [11] (2025)	60 common and 104 complex staged cases	Advanced models exceeded 90% in some common-case settings; Claude 3.7 reached 83.3% at final complex-case stage.	Curated sequential cases; model- and stage-specific.
Omar [16] (2025)	300 seeded vignettes; 5,400 model outputs	50.0%–82.7% hallucination under default adversarial configurations across specified models/formats.	False-fact stress testing; no patient-harm measurement.
Chang [17] (2025)	376 red-team prompts; 1,504 outputs	303/1,504 (20.1%) rated inappropriate; 156/303 inappropriate outputs included hallucinations.	Adversarial prompts; output appropriateness differs from clinical event rates.
Xu [29] (2025)	48 short vignettes; eight LLMs	Mean diagnosis 89.84% unprompted vs 91.67% prompted; triage 76.82% vs 86.20%.	Prompt condition used 47 of 48 benchmark cases with gold-standard answers.
Rao [12] (2026)	29 sequential vignettes; 21 LLMs	Performance varied by reasoning stage; differential construction was weaker than final diagnosis.	Benchmark setting; not continuous patient management.
Pendse [20] (2026)	97 psychiatric cases; demographic permutations	15,346/18,527 outputs (82.8%) had ≥1 ground-truth diagnosis; 13,017/18,527 (70.3%) had ≥1 non-reference-set label.	Repeated vignette audit; clinically unverified added labels.
van Kessel [18] (2026)	12,197 valid outputs; two disease vignettes	Guideline omissions up to 97% (DeepSeek-V3) and 46% (GPT-4.1); hallucinated guidelines up to 9%.	Two diseases; specific zero-shot guideline task.

Study	Study design, population, and unit of analysis	Reported task-specific outcome (as published)	Critical limitation / clinical applicability
Sheppert [21] (2026)	Nine paired vignettes; eight LLMs and residents	Implausible anchor placed first: 40/72 (55.6%) model vs 17/80 (21.2%) resident responses.	Forced-choice anchoring paradigm.
Magdziarz Ibrahim-El-Nur [22] (2026)	30 matched English/Polish cases; six LLMs	Diagnostic workup and differential quality varied by model and language.	No Arabic testing; translation-controlled task.
Brodeur [30] (2026)	Five reasoning experiments; ED second-opinion study	LLM exceeded tested physician baselines in reported reasoning tasks.	Distinct experimental tasks; prospective clinical validation remains necessary.

The study-specific units of analysis include cases, physicians, and repeated model outputs; cross-study arithmetic comparisons are inappropriate

4.2. Standalone diagnostic accuracy and sequential reasoning

In 52 complex published cases Hirosawa et al. [5] found 31/52 (60%) correct first-ranked diagnoses, 42/52 (81%) top five coverage and 43/52 (83%) top ten coverage. These are nested not independent results. Kanjee et al. [6] had a correct top diagnosis in 27 of 70 cases (39%) and it was included anywhere in the differential in 45 of 70 (64%). Ríos-Hoyo et al. [7] top one accuracy was 22% and differential inclusion 68% across 75 complex published cases. Lower or higher proportions cannot be ranked as model comparisons since reference cases, output length and scoring differ.

The stage and information available at inference also affected apparent performance. In 36 previously unpublished cases Feldman et al. [10] reported top-25 correct diagnosis counts without lab info for DXplain (20/36), GPT-4 (15/36) and Gemini (14/36); with lab info counts were 26/36, 23/36 and 21/36. A top-25 match should not be confused with a safe primary diagnosis. Cabral et al. [8], Hager et al. [9], Dinc et al. [11] and Rao et al. [12] further showed how sensitive clinical reasoning assessment was to how information, model outputs and reference standards were structured.

Xu et al [29] tested eight models on 48 short clinical vignettes and reported mean diagnostic accuracy without the evaluated prompt was 89.84% vs 91.67% with it; triage accuracy reported was 76.82% vs 86.20%. The prompted condition contained answers from 47 of 48 benchmark cases (potentially answer revealing task support); these numbers should not be presented as independent validation of generalizable prompting benefits. Over-triage increases also show why a gain in one metric can coexist with a clinical trade-off.

4.3. Clinician–LLM collaboration

Goh et al [13] randomized 50 physicians to GPT-4 plus usual diagnostic resources or usual resources alone: median case-level diagnostic reasoning scores were 76% (IQR, 66%–87%) versus 74% (IQR, 63%–84%) (adjusted difference 2 percentage points; 95% CI −4 to 8; P = 0.60). An exploratory evaluation of GPT-4 in isolation had median score 92% (IQR 82%–97%) and 16 percentage point advantage over conventional resource comparison (95% CI 2–30; P = 0.03). The model-alone result is not evidence that giving physicians access to the model improved performance; nor does it establish patient benefit.

Using 302 selected published cases McDuff et al. [15] found 51.7% clinician top-ten differential coverage with AMIE assistance vs 36.1% without AMIE; search resources were 44.4%. AMIE alone reached 59.1% top ten coverage vs 33.6% among unassisted clinicians in a separate comparison. Tu et al. [14] tested text-only simulated consultations (159 scenarios and 20 PCPs) on AMIE ratings on many specialist and patient-actor dimensions using

Brodeur et al.'s strong task-specific reasoning results and an emergency-department second opinion comparison later. These studies differ in model, workflow, human comparator and outcome so no pooled conclusion about clinician augmentation in routine practice can be drawn from them.

4.4 Adversarial safety, triage and guideline reliability

In an adversarial false-information study Omar et al [16] analysed 5400 outputs from seeded clinical vignettes and found default-configuration hallucination rates of 50.0% to 82.7% for specified model/format pairs. Overall rate under default prompting was 65.9%; with mitigation prompt it was 44.2%. This is under deliberately corrupted inputs not background rates of routine erroneous output or patient injury.

Chang et al. [17] analysed 376 red-team prompts for four model configurations (1,504 outputs) and found 303 (20.1%) inappropriate responses; 156 of those 303 inappropriate responses (51.3%) contained hallucinations. In a guideline referencing study hypercholesterolemia and type 2 diabetes van Kessel et al. [18] analysed 12,197 valid outputs: guideline omissions up to 97% for DeepSeek-V3 and 46% for GPT-4.1 in tested disease conditions and hallucinated guideline references up to about 9%. Units of experiment and scoring thresholds vary between studies. None gives incidence of preventable patient harm from deploying LLMs.

4.5 Demographic fairness, diagnostic anchoring and language: Counterfactual evaluations showed differences in generated content when demographic descriptions were different [19,20]. In a psychiatric diagnostic audit Pendse et al. [20] found at least one reference-consistent diagnosis for 15,346/18,527 (82.8%) GPT-4o outputs but also at least one diagnosis outside the vignette reference set in 13,017/18,527 (70.3%). These two outcomes can coexist in the same response so wide differential coverage is not enough safety metric. Whether a demographic difference is appropriate disease prevalence adjustment or stereotyping needs to be clinically assessed.

Sheppert et al. [21] found that an implausible diagnosis proposed by the patient was ranked first in 40/72 LLM responses (55.6%) compared to 17/80 resident responses (21.2%) under forced-choice vignette conditions; these are task-specific anchoring responses and not directly measured wrong diagnoses in routine encounters. Magdziarz Ibrahim-El Nur et al. [22] showed language-/model-specific differences in translated English/Polish vignettes but neither experiment nor other studies here report diagnostic error rates in Arabic-language consultations.

4.6 Applicability of the evidence

External validity is limited by selected and sometimes public case material, benchmark-answer exposure (model/version drift), prompt sensitivity, nonuniform reference standards (top-k cutoff), different top-k cutoffs, simulated contexts etc. none of Tables 1A–1B quantifies prospective preventable diagnostic injury; Table 2 separates study specific numerical results from invalid extrapolations; Table 3 lists proposed preimplementation evaluation domains and is not validated instrument. Future updates should include independent clinical adjudication and outcome definitions aligned with intended real world use.

Table 2. Quantitative results and appropriate interpretation.

Evidence domain / endpoint	Original study-specific finding	Limits of interpretation / clinical applicability
Case-based accuracy	Hirosawa rank 1: 31/52 (60%); top 10: 43/52 (83%) [5].	Top-k matches on selected published cases do not quantify prospective clinical accuracy.
Clinician trial	Physician score 76% vs 74%; adjusted Δ2 points, 95% CI −4 to 8; P=.60 [13].	A non-significant result is not proof of equivalence or generalized ineffectiveness.
Assisted differential	AMIE-assisted 51.7% vs unassisted 36.1% top-10 [15].	Task-specific clinician assistance; published-case evaluation, not patient outcomes.

Evidence domain / endpoint	Original study-specific finding	Limits of interpretation / clinical applicability
Triage/prompting	Xu triage 76.82% vs 86.20%; prompted design used 47 reference-answered cases [29].	Answer-revealing support and over-triage prevent straightforward deployment inference.
Adversarial reliability	Omar 50.0%–82.7% hallucination in tested default model/case-format conditions [16].	Stress test under seeded false facts; not baseline patient-harm rate.
Guideline omissions	Up to 97% DeepSeek-V3 and 46% GPT-4.1 omissions in two disease tasks [18].	Conditional task- and model-specific omissions.
Fairness and overgeneration	Pendse: 82.8% ≥1 correct label; 70.3% ≥1 non-reference-set label [20].	A response can contain both true and ungrounded labels; audit is psychiatric-vignette based.
Anchoring	40/72 models vs 17/80 residents ranked implausible anchor first [21].	Forced-choice experimental effect; not comparative routine error risk.

Table 3. Proposed domains for evaluating LLM diagnostic decision support before clinical use (conceptual framework, not a validated checklist).

Preimplementation domain	Prespecified evaluation question	Minimum evidence / safeguards
Technical validity	Prespecified clinical indication, population, model/version, prompt, and independent test set.	Clinical task representativeness; data leakage; repeated runs; reproducibility.
Diagnostic utility	Top-one and clinically prioritized differential; missed time-critical disease; next-test quality.	Prospective clinician–LLM comparator trials; prespecified adjudication.
Safety and oversight	Hallucinated evidence/guidelines, triage failures, escalation, abstention, and audit trail.	Severity-weighted safety endpoints and real clinical consequence.
Equity and language	Subgroups by age, sex, language, literacy, and context; clinically warranted prevalence effects.	Representative Arabic–English testing and subgroup calibration.
Operational readiness	Integration, governance, access control, privacy, clinician responsibility, change management.	Prospective outcome monitoring, drift detection, and incident review.

Interpretation of the evidence should follow an endpoint hierarchy: (a) externally adjudicated, prospective patient safety and diagnostic outcomes; (b) controlled physician–AI workflow outcomes; (c) independently assessed sequential clinical decisions; and (d) retrospective or vignette-level concordance and adversarial stress tests. This hierarchy is a proposed framework for interpretation, not a validated grading system.

5. DISCUSSION

The main takeaway is not that LLMs have a single "diagnostic accuracy" level but that reported performance varies with clinical task, benchmark, permitted information, model version and scoring standard. First ranked diagnosis

vs inclusion anywhere in differential vs top-25 coverage vs next test quality vs safe triage are different constructs. A correct disease buried in a long list does not establish useful prioritization of emergency diagnoses. Future primary studies should prespecify clinically meaningful error severity and representative comparator (not just concordance with reference label) [5–12]. Clinician plus LLM combination needs direct evaluation: the exploratory standalone GPT-4 result in Goh et al. [13] did not translate into statistically significant improvement when randomized physicians had access to model; AMIE-assisted differential coverage and simulated conversation ratings [14,15] were better under different interfaces and evaluation protocols. These results cannot support ranking tools cross study because clinician uptake training baseline resources case mix endpoints varied. Randomized comparisons of workflows and patient safety outcomes are needed. Takita et al. broad synthesis [23] aggregated 83 studies and estimated overall generative-AI diagnostic accuracy at 52.1%, not significantly different from physicians as a whole but lower than expert physicians in the comparison reported. This estimate should not be summed with top-k and task-specific percentages presented here (cases, endpoints, inclusion approach, period differ), because they are different. What makes this narrative review distinctive is not an accuracy estimate competing against these percentages but cross domain assessment of how performance claims relate to deployment safety, algorithmic fairness and practical clinical oversight. Stress testing shows failure modes rather than rates of harm directly observed. The seeded false fact and red team studies [16,17] show susceptibility to misinformation supplied externally and inappropriate clinical text; guideline omission study [18] shows variability even when models are asked for established guidance. Clinical risk may also arise from omission, delayed escalation, misplaced diagnostic confidence or recommendations unsupported by patient information. Safety monitoring therefore needs to record severity weighted errors (missed serious diagnoses), unnecessary procedures (delays) and clinician acceptance or override decisions (not only exact match accuracy). Fairness assessment requires more than equal raw percentages. Shifts under demographic substitutions may be due to harmful stereotyping, appropriate differences in prior probability or problems with the reference diagnosis [19,20]. Forced-choice anchor experiment [21] shows susceptibility to irrelevant patient suggestions but not absolute frequency of anchoring in routine clinical documentation. Translation controlled English–Polish evidence [22] suggests a plausible need for multilingual evaluation without showing impairment in Arabic. A Saudi agenda for validation should include Arabic–English code-switching; local terminology; disease patterns by region; emergency pathways; stratified adjudication. Clinical implementation should be staged and governed (use case, intended user, setting, model version allowed data human verification step escalation rules accountable clinical owner). Independent local validation should include representative cases time critical misses' uncertainty/abstention behaviour hallucinated references privacy protection incident reporting repeat testing after model or prompt updates. WHO governance guidance [24] and study-specific reporting standards such as STARD-AI TRIPOD-LLM DECIDE-AI [31–33] are useful where applicable but not substitutes for prospective safety evaluation or regulatory requirements. A prospective validation pathway is proposed: (1) register a clinical indication and protocol; (2) lock model version and prompts; (3) validate against independent local retrospective cases; (4) perform silent-mode prospective evaluation; (5) compare clinician-assisted versus usual-care decisions using predefined serious-miss and unsafe-triage endpoints; and (6) monitor adverse events, subgroup gaps, workload, overrides and software updates. Where clinical risk is high, independent safety oversight and escalation should precede deployment. This review has limitations. (1) The 20-studies source set is author curated without reproducible search log or excluded-record archive so selection and publication biases remain possible. (2) Duplicate extraction was not performed, no formal risk of bias assessment, certainty grading nor independent meta-analysis were carried out. (3) Case sources, model versions, reference standards, prompts, denominators and outcomes vary across reports. (4) Benchmark exposure and answer revealing prompts may inflate some results. (5) Adversarial output rates and differential list matches are surrogates not observed patient harm or net clinical benefit. (6) Arabic language deployment, multi-site generalizability, subgroup equity, operational cost and longitudinal monitoring were not represented adequately. These limitations should moderate causal and clinical implementation claims.

6. CONCLUSIONS.

Large language models can be task-specific diagnosticians but have not shown reliable autonomous diagnostic safety or improved patient outcomes in routine clinical settings. Standalone benchmark performance cannot be extrapolated to physician assistance effectiveness and stress test error frequencies cannot be read as patient harm

incidence. The right near-term research priority is prospective assessment of well-defined workflows supervised by clinicians (urgent condition detection, next test/triage decisions, independent clinical adjudication, longitudinal harms, subgroup fairness, Arabic–English usability and post deployment model drift).

Author and Ethics Declarations

Author Contributions: Conceptualization, Awadalla Abdelwahid and Fatima Habib Allah Salama; methodology, Awadalla Abdelwahid, Abdelazim Ahmed Idris, and Dalia Omar; literature searching and evidence identification, Rana Fath Elrhman, Dalya I. A. Mohamed, Dina Mahmoud Fathi, and Mayada Farah Obaid; evidence evaluation and critical synthesis, Awadalla Abdelwahid, Fatima Habib Allah Salama, Hager Ahmed Elhabo Mohamed, and Khalid Yousif Fakhri; writing—original draft preparation, Awadalla Abdelwahid, Dalia Omar, Rana Fath Elrhman, and Dina Mahmoud Fathi; writing—review and editing, all authors; supervision, Awadalla Abdelwahid and Abdelazim Ahmed Idris; project administration, Awadalla Abdelwahid.

All authors critically reviewed the manuscript for important intellectual content, approved the final version for publication, and agreed to be accountable for their respective contributions to the work.

Funding: Not funded

Institutional review board statement: Not applicable; this article is a narrative review of previously published literature and involved no new recruitment or primary patient-level data collection.

Informed consent statement: Not applicable; no new individual participant information is reported.

Data availability statement: No new primary patient-level data were collected, and no new pooled dataset or original meta-analysis was generated. All discussed evidence is available through the cited publications.

Conflicts of interest: All Authors declare no conflict of interest

Use of generative AI: Not used. All authors have independently reviewed and verified the references, citations, scientific interpretations, and manuscript content. All authors have reviewed and approved the final version of the manuscript and accept full responsibility for its accuracy, integrity, and scientific content in accordance with the target journal's editorial policies.

REFERENCE

[1] National Academies of Sciences, Engineering, and Medicine. Improving Diagnosis in Health Care. Washington (DC): National Academies Press; 2015. https://doi.org/10.17226/21794

[2] Sutton RT, Pincock D, Baumgart DC, et al. An overview of clinical decision support systems: benefits, risks, and strategies for success. NPJ Digit Med. 2020;3:17. https://doi.org/10.1038/s41746-020-0221-y

[3] Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. Nature. 2023;620:172–180. https://doi.org/10.1038/s41586-023-06291-2

[4] Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health. 2023;2:e0000198. https://doi.org/10.1371/journal.pdig.0000198

[5] Hirosawa T, Kawamura R, Harada Y, et al. ChatGPT-generated differential diagnosis lists for complex case-derived clinical vignettes: diagnostic accuracy evaluation. JMIR Med Inform. 2023;11:e48808. https://doi.org/10.2196/48808

[6] Kanjee Z, Crowe B, Rodman A. Accuracy of a generative artificial intelligence model in a complex diagnostic challenge. JAMA. 2023;330:78–80. https://doi.org/10.1001/jama.2023.8288

[7] Ríos-Hoyo A, Shan NL, Li A, Pearson AT, Pusztai L, Howard FM. Evaluation of large language models as a diagnostic aid for complex medical cases. Front Med (Lausanne). 2024;11:1380148. https://doi.org/10.3389/fmed.2024.1380148

[8] Cabral S, Restrepo D, Kanjee Z, et al. Clinical reasoning of a generative artificial intelligence model compared with physicians. JAMA Intern Med. 2024;184:581–583. https://doi.org/10.1001/jamainternmed.2024.0295

[9] Hager P, Jungmann F, Holland R, et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. Nat Med. 2024;30:2613–2622. https://doi.org/10.1038/s41591-024-03097-1

[10] Feldman MJ, Hoffer EP, Conley JJ, et al. Dedicated AI expert system vs generative AI with large language model for clinical diagnoses. JAMA Netw Open. 2025;8:e2512994. https://doi.org/10.1001/jamanetworkopen.2025.12994

[11] Dinc MT, Bardak AE, Bahar F, Noronha C. Comparative analysis of large language models in clinical diagnosis: performance evaluation across common and complex medical cases. JAMIA Open. 2025;8:ooaf055. https://doi.org/10.1093/jamiaopen/ooaf055
[12] Rao AS, Esmail KP, Lee RS, et al. Large language model performance and clinical reasoning tasks. JAMA Netw Open. 2026;9:e264003. https://doi.org/10.1001/jamanetworkopen.2026.4003
[13] Goh E, Gallo R, Hom J, et al. Large language model influence on diagnostic reasoning: a randomized clinical trial. JAMA Netw Open. 2024;7:e2440969. https://doi.org/10.1001/jamanetworkopen.2024.40969
[14] Tu T, Schaekermann M, Palepu A, et al. Towards conversational diagnostic artificial intelligence. Nature. 2025;642:442–450. https://doi.org/10.1038/s41586-025-08866-7
[15] McDuff D, Schaekermann M, Tu T, et al. Towards accurate differential diagnosis with large language models. Nature. 2025;642:451–457. https://doi.org/10.1038/s41586-025-08869-4
[16] Omar M, Sorin V, Collins JD, et al. Multi-model assurance analysis showing large language models are highly vulnerable to adversarial hallucination attacks during clinical decision support. Commun Med (Lond). 2025;5:330. https://doi.org/10.1038/s43856-025-01021-3
[17] Chang CT, Farah H, Gui H, et al. Red teaming ChatGPT in medicine to yield real-world insights on model behavior. NPJ Digit Med. 2025;8:149. https://doi.org/10.1038/s41746-025-01542-0
[18] van Kessel R, Anderson M, McMillan B, et al. Omission and hallucination prevalence of clinical guidelines in diagnostic large language model outputs. BMJ Health Care Inform. 2026;33:e101959. https://doi.org/10.1136/bmjhci-2025-101959
[19] Zack T, Lehman E, Suzgun M, et al. Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study. Lancet Digit Health. 2024;6:e12–e22. https://doi.org/10.1016/S2589-7500(23)00225-X
[20] Pendse SR, Jain M, Kumar N, De Choudhury M. Implicit gender, racial, and ethnic biases in large language models: an audit study of automated psychiatric diagnoses. Mayo Clin Proc Digit Health. 2026;4:100344. https://doi.org/10.1016/j.mcpdig.2026.100344
[21] Sheppert A, Shen C, Geissal E, Riley C, Riley S. Large language models exhibit greater diagnostic anchoring than physicians in a forced-choice vignette study. Int J Med Inform. 2026;219:106550. https://doi.org/10.1016/j.ijmedinf.2026.106550
[22] Magdziarz Ibrahim-El-Nur J, Kaczmarek W, Winiarska W, Kuś A, Łoś M. Language-specific differences in large language model diagnostic reasoning: a translation-controlled clinical vignette study. J Clin Med. 2026;15:4082. https://doi.org/10.3390/jcm15114082
[23] Takita H, Kabata D, Walston SL, et al. A systematic review and meta-analysis of diagnostic performance comparison between generative AI and physicians. NPJ Digit Med. 2025;8:175. https://doi.org/10.1038/s41746-025-01543-z
[24] World Health Organization. Ethics and governance of artificial intelligence for health: guidance on large multi-modal models. Geneva: World Health Organization; 2024. ISBN: 978-92-4-008475-9. https://www.who.int/publications/b/70584
[25] Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021;372:n71. https://doi.org/10.1136/bmj.n71
[26] Baethge C, Goldbeck-Wood S, Mertens S. SANRA—a scale for the quality assessment of narrative review articles. Res Integr Peer Rev. 2019;4:5. https://doi.org/10.1186/s41073-019-0064-8
[27] Whiting PF, Rutjes AWS, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med. 2011;155:529–536. https://doi.org/10.7326/0003-4819-155-8-201110180-00009
[28] Moons KGM, Damen JAA, Kaul T, et al. PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. BMJ. 2025;388:e082505. https://doi.org/10.1136/bmj-2024-082505
[29] Xu L, Zhao W, Huang X. Diagnosis and triage performance of contemporary large language models on short clinical vignettes. J Med Syst. 2025;49:141. https://doi.org/10.1007/s10916-025-02284-y
[30] Brodeur PG, Buckley TA, Kanjee Z, et al. Performance of a large language model on the reasoning tasks of a physician. Science. 2026;392(6797):524–527. https://doi.org/10.1126/science.adz4433
[31] Sounderajah V, Guni A, Liu X, et al. The STARD-AI reporting guideline for diagnostic accuracy studies using artificial intelligence. Nat Med. 2025;31:3283–3289. https://doi.org/10.1038/s41591-025-03953-8
[32] Gallifant J, Afshar M, Ameen S, et al. The TRIPOD-LLM reporting guideline for studies using large language models. Nat Med. 2025;31:60–69. https://doi.org/10.1038/s41591-024-03425-5

[33] Vasey B, Nagendran M, Campbell B, et al. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. BMJ. 2022;377:e070904. https://doi.org/10.1136/bmj-2022-070904