

Supplementary Note 1: PRISMA 2020 checklist

Section and Topic	Item #	Checklist item	Location where item is reported
TITLE			
Title	1	Identify the report as a systematic review.	Title
ABSTRACT			
Abstract	2	See the PRISMA 2020 for Abstracts checklist.	End of this document
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of existing knowledge.	First, second paragraph
Objectives	4	Provide an explicit statement of the objective(s) or question(s) the review addresses.	Third paragraph
METHODS			
Eligibility criteria	5	Specify the inclusion and exclusion criteria for the review and how studies were grouped for the syntheses.	Search strategy and eligibility criteria
Information sources	6	Specify all databases, registers, websites, organisations, reference lists and other sources searched or consulted to identify studies. Specify the date when each source was last searched or consulted.	Search strategy and eligibility criteria
Search strategy	7	Present the full search strategies for all databases, registers and websites, including any filters and limits used.	Supplementary Note 2
Selection process	8	Specify the methods used to decide whether a study met the inclusion criteria of the review, including how many reviewers screened each record and each report retrieved, whether they worked independently, and if applicable, details of automation tools used in the process.	Search strategy and eligibility criteria
Data collection process	9	Specify the methods used to collect data from reports, including how many reviewers collected data from each report, whether they worked independently, any processes for obtaining or confirming data from study investigators, and if applicable, details of automation tools used in the process.	Data extraction
Data items	10a	List and define all outcomes for which data were sought. Specify whether all results that were compatible with each outcome domain in each study were sought (e.g. for all measures, time points, analyses), and if not, the methods used to decide which results to collect.	Data extraction

Section and Topic	Item #	Checklist item	Location where item is reported
	10b	List and define all other variables for which data were sought (e.g. participant and intervention characteristics, funding sources). Describe any assumptions made about any missing or unclear information.	Data extraction
Study risk of bias assessment	11	Specify the methods used to assess risk of bias in the included studies, including details of the tool(s) used, how many reviewers assessed each study and whether they worked independently, and if applicable, details of automation tools used in the process.	Quality assessment, supplementary note 3
Effect measures	12	Specify for each outcome the effect measure(s) (e.g. risk ratio, mean difference) used in the synthesis or presentation of results.	Data analysis
Synthesis methods	13a	Describe the processes used to decide which studies were eligible for each synthesis (e.g. tabulating the study intervention characteristics and comparing against the planned groups for each synthesis (item #5)).	Data analysis
	13b	Describe any methods required to prepare the data for presentation or synthesis, such as handling of missing summary statistics, or data conversions.	Data analysis
	13c	Describe any methods used to tabulate or visually display results of individual studies and syntheses.	Data analysis
	13d	Describe any methods used to synthesize results and provide a rationale for the choice(s). If meta-analysis was performed, describe the model(s), method(s) to identify the presence and extent of statistical heterogeneity, and software package(s) used.	Data analysis
	13e	Describe any methods used to explore possible causes of heterogeneity among study results (e.g. subgroup analysis, meta-regression).	Data analysis
	13f	Describe any sensitivity analyses conducted to assess robustness of the synthesized results.	Data analysis
Reporting bias assessment	14	Describe any methods used to assess risk of bias due to missing results in a synthesis (arising from reporting biases).	Quality assessment, data analysis
Certainty assessment	15	Describe any methods used to assess certainty (or confidence) in the body of evidence for an outcome.	Quality assessment, data analysis
RESULTS			
Study selection	16a	Describe the results of the search and selection process, from the number of records identified in the search to the number of studies included in	Study selection, figure 1

Section and Topic	Item #	Checklist item	Location where item is reported
		the review, ideally using a flow diagram.	
	16b	Cite studies that might appear to meet the inclusion criteria, but which were excluded, and explain why they were excluded.	Figure 1
Study characteristics	17	Cite each included study and present its characteristics.	Study characteristics
Risk of bias in studies	18	Present assessments of risk of bias for each included study.	Quality assessment
Results of individual studies	19	For all outcomes, present, for each study: (a) summary statistics for each group (where appropriate) and (b) an effect estimate and its precision (e.g. confidence/credible interval), ideally using structured tables or plots.	Table 2, Figures 2 to 4
Results of syntheses	20a	For each synthesis, briefly summarise the characteristics and risk of bias among contributing studies.	Table 1 and supplementary figures 2 and 3
	20b	Present results of all statistical syntheses conducted. If meta-analysis was done, present for each the summary estimate and its precision (e.g. confidence/credible interval) and measures of statistical heterogeneity. If comparing groups, describe the direction of the effect.	Table 2, Figures 2 to 4
	20c	Present results of all investigations of possible causes of heterogeneity among study results.	Table 2, Figure 3, Table S1 and S2
	20d	Present results of all sensitivity analyses conducted to assess the robustness of the synthesized results.	Table S3 and S4
Reporting biases	21	Present assessments of risk of bias due to missing results (arising from reporting biases) for each synthesis assessed.	Table S2, S3 and S4
Certainty of evidence	22	Present assessments of certainty (or confidence) in the body of evidence for each outcome assessed.	Table 2, Figures 2 to 4
DISCUSSION			
Discussion	23a	Provide a general interpretation of the results in the context of other evidence.	First paragraph

Section and Topic	Item #	Checklist item	Location where item is reported
	23b	Discuss any limitations of the evidence included in the review.	3 rd to 6 th paragraphs in Discussion
	23c	Discuss any limitations of the review processes used.	Study limitations section
	23d	Discuss implications of the results for practice, policy, and future research.	13 th paragraph in Discussion
OTHER INFORMATION			
Registration and protocol	24a	Provide registration information for the review, including register name and registration number, or state that the review was not registered.	First paragraph in the method
	24b	Indicate where the review protocol can be accessed, or state that a protocol was not prepared.	First paragraph in the method
	24c	Describe and explain any amendments to information provided at registration or in the protocol.	NA
Support	25	Describe sources of financial or non-financial support for the review, and the role of the funders or sponsors in the review.	Acknowledgement
Competing interests	26	Declare any competing interests of review authors.	Competing interest
Availability of data, code and other materials	27	Report which of the following are publicly available and where they can be found: template data collection forms; data extracted from included studies; data used for all analyses; analytic code; any other materials used in the review.	Data availability and code availability

PRISMA 2020 for Abstracts Checklist

Section and Topic	Item #	Checklist item	Reported (Yes/No)
TITLE			
Title	1	Identify the report as a systematic review.	Yes

Section and Topic	Item #	Checklist item	Reported (Yes/No)
BACKGROUND			
Objectives	2	Provide an explicit statement of the main objective(s) or question(s) the review addresses.	Yes
METHODS			
Eligibility criteria	3	Specify the inclusion and exclusion criteria for the review.	Yes
Information sources	4	Specify the information sources (e.g. databases, registers) used to identify studies and the date when each was last searched.	No
Risk of bias	5	Specify the methods used to assess risk of bias in the included studies.	No
Synthesis of results	6	Specify the methods used to present and synthesise results.	Yes
RESULTS			
Included studies	7	Give the total number of included studies and participants and summarise relevant characteristics of studies.	Yes
Synthesis of results	8	Present results for main outcomes, preferably indicating the number of included studies and participants for each. If meta-analysis was done, report the summary estimate and confidence/credible interval. If comparing groups, indicate the direction of the effect (i.e. which group is favoured).	Yes
DISCUSSION			
Limitations of evidence	9	Provide a brief summary of the limitations of the evidence included in the review (e.g. study risk of bias, inconsistency and imprecision).	Yes
Interpretation	10	Provide a general interpretation of the results and important implications.	Yes
OTHERS			
Funding	11	Specify the primary source of funding for the review.	No
Registration	12	Provide the register name and registration number	No

Supplementary Note 2: Search strategy

1. PubMed (1972 results, 10th September 2025)

("large language model*"[Title/Abstract] OR "LLM"[Title/Abstract] OR "generative artificial intelligence"[Title/Abstract] OR "generative AI"[Title/Abstract] OR "generative pre-trained transformer*"[Title/Abstract] OR "chatbot*"[Title/Abstract] OR "GPT"[Title/Abstract] OR "ChatGPT"[Title/Abstract] OR "Claude"[Title/Abstract] OR "Llama"[Title/Abstract] OR "PaLM"[Title/Abstract] OR "Gemini"[Title/Abstract])
AND ("Diagnosis, Computer Assisted"[MeSH Terms] OR "Triage"[MeSH Terms] OR "diagnos*"[Title/Abstract] OR "triag*"[Title/Abstract])
AND ("accurac*" [Title/Abstract] OR "correct*" [Title/Abstract] OR "agreement" [Title/Abstract] OR "sensitivity" [Title/Abstract] OR "specificity" [Title/Abstract] OR "precision" [Title/Abstract] OR "recall" [Title/Abstract] OR "F1" [Title/Abstract] OR "AUC" [Title/Abstract])
NOT ("review"[Publication Type] OR "meta-analysis"[Publication Type] OR "systematic review"[Publication Type])
AND (2020:2025[pdat])

2. Embase (3786 results, 10th September 2025)

('large language model*':ab,ti OR 'LLM':ab,ti OR 'generative artificial intelligence':ab,ti OR 'generative AI':ab,ti OR 'generative pre-trained transformer*':ab,ti OR 'chatbot*':ab,ti OR 'GPT':ab,ti OR 'ChatGPT':ab,ti OR 'Claude':ab,ti OR 'Llama':ab,ti OR 'PaLM':ab,ti OR 'Gemini':ab,ti)
AND ('diagnosis'/exp OR 'diagnos*':ab,ti OR 'triage':ab,ti)
AND ('accurac*':ab,ti OR 'correct*':ab,ti OR 'agreement':ab,ti OR 'sensitivity':ab,ti OR 'specificity':ab,ti OR 'precision':ab,ti OR 'recall':ab,ti OR 'F1':ab,ti OR 'AUC':ab,ti)
NOT ('review':it OR 'systematic review':it OR 'meta-analysis':it)
AND [2020-2025]/py

3. IEEE Xplore (480 results, 10th September 2025)

((("All Metadata": "large language model*") OR ("All Metadata": LLM) OR ("All Metadata": "generative artificial intelligence") OR ("All Metadata": "generative AI") OR ("All Metadata": "generative pre-trained transformer*") OR ("All Metadata": "chatbot*") OR ("All Metadata": GPT) OR ("All Metadata": ChatGPT) OR ("All Metadata": Claude) OR ("All Metadata": Llama) OR ("All Metadata": PaLM) OR ("All Metadata": Gemini)) AND ((("All Metadata": "diagnos*") OR ("All Metadata": "triag*")) AND ((("All Metadata": "accurac*") OR ("All Metadata": "correct*") OR ("All Metadata": "agreement") OR ("All Metadata": "sensitivity") OR ("All Metadata": "specificity") OR ("All Metadata": "precision") OR ("All Metadata": "recall") OR ("All Metadata": "F1") OR ("All Metadata": "AUC")) NOT (("Publication Title": review) OR ("Publication Title": systematic review) OR ("Publication Title": meta-analysis)))
Limit to year = "2020-2025"
Limit to Content Type = "Journal"

4. Cochrane Library (1187 results, 10th September 2025)

[mh "Large Language Models"] OR [mh "Generative Artificial Intelligence"] OR [mh "Artificial Intelligence"] OR "large language model*":ti,ab,kw OR LLM:ti,ab,kw OR "generative artificial

intelligence":ti,ab,kw OR "generative AI":ti,ab,kw OR "generative pre-trained transformer*":ti,ab,kw OR chatbot*:ti,ab,kw OR GPT:ti,ab,kw OR ChatGPT:ti,ab,kw OR Claude:ti,ab,kw OR Llama:ti,ab,kw OR PaLM:ti,ab,kw OR Gemini:ti,ab,kw)
AND ([mh "Diagnosis, Computer-Assisted"] OR [mh "Triage"] OR diagnos*:ti,ab,kw OR triag*:ti,ab,kw)
AND ([mh "Sensitivity and Specificity"] OR accurac*:ti,ab,kw OR correct*:ti,ab,kw OR agreement:ti,ab,kw OR sensitivity:ti,ab,kw OR specificity:ti,ab,kw OR precision:ti,ab,kw OR recall:ti,ab,kw OR F1:ti,ab,kw OR AUC:ti,ab,kw)
NOT (review:pt OR "systematic review":pt OR "meta analysis":pt)
Limit to Content type= "Trials"
Limit to publication year Between 2020 and 2025

5. Epistemonikos (458 results, 10th September 2025)

("large language model*" OR LLM OR "generative artificial intelligence" OR "generative AI" OR "generative pre-trained transformer*" OR chatbot* OR GPT OR ChatGPT OR Claude OR Llama OR PaLM OR Gemini)
AND (diagnos* OR triag*)
AND (accurac* OR correct* OR agreement OR sensitivity OR specificity OR precision OR recall OR F1 OR AUC)
NOT (review OR "systematic review" OR "meta-analysis")
Filter Publication year from 2020 to 2025

6. CINAHL (EBSCO) (893 results, 10th September 2025)

(MH "Artificial Intelligence, Generative+" OR MH "Artificial Intelligence+" OR MH "Chatbot" OR "large language model*":TI,AB OR LLM:TI,AB OR "generative pre-trained transformer*":TI,AB OR GPT:TI,AB OR ChatGPT:TI,AB OR Claude:TI,AB OR Llama:TI,AB OR PaLM:TI,AB OR Gemini:TI,AB)
AND (MH "Diagnosis, Computer Assisted+" OR MH "Triage" OR diagnos*:TI,AB OR triag*:TI,AB)
AND (MH "Sensitivity and Specificity" OR accurac*:TI,AB OR correct*:TI,AB OR agreement:TI,AB OR sensitivity:TI,AB OR specificity:TI,AB OR precision:TI,AB OR recall:TI,AB OR F1:TI,AB OR AUC:TI,AB)
NOT (PT Review OR PT "Systematic Review" OR PT "Meta Analysis")
Limiters: Published Date: 20200101–20251231

7. PsycINFO (Ovid) (1617 results, 10th September 2025)

(exp Artificial Intelligence/ or Large Language Models/ or Generative Artificial Intelligence/ or "large language model*".ti,ab. or LLM.ti,ab. or "generative pre-trained transformer*".ti,ab.)
AND (Diagnosis, Computer-Assisted/ or Triage/ or diagnos*.ti,ab. or triag*.ti,ab.)
AND (Sensitivity.mp. and Specificity/)
NOT (review.pt. OR "systematic review".pt. OR "meta analysis".pt.)
limit to yr="2020 - 2025"

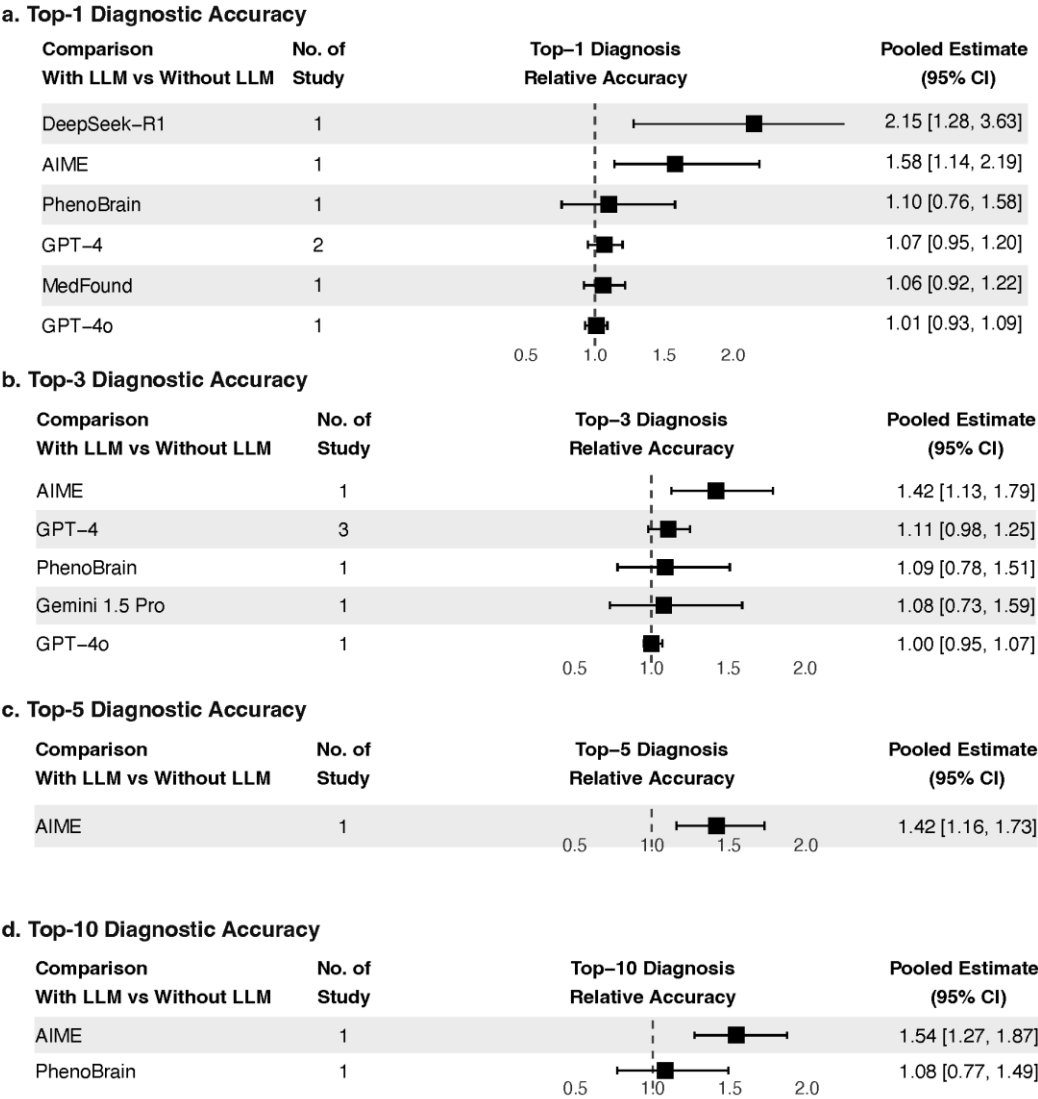
Supplementary Note 3: The modified version of the Quality of Diagnostic Accuracy Studies 2 (QUADAS-2)

Domains	Signal questions	Explanation
Risk of bias		
Participant selection	Was a consecutive or random sample of cases enrolled?	Yes - RCTs and cohort studies (prospective or retrospective) with unenriched (consecutive or random) sampling. Unclear - If not stated. No - Other studies.
	Did the study avoid inappropriate case exclusions?	Yes - If inappropriate exclusions were avoided. Unclear - If not clearly reported. No - Other studies.
Index tests	Were the index test results interpreted without knowledge of the results of the reference standard?	Yes - Require clear statement of blinding, or clear temporal relationships where the index test occurred before the reference standard. Unclear - Not clearly state whether blinding was implemented or not provide enough detail to determine this. No - The index test was conducted with knowledge of the reference standard results.
	The LLM and versions used (e.g., GPT-4), Date of inquiry (e.g., the cases was posed to ChatGPT on March 10th, 2023)	Yes - The LLMs used and their versions were reported. Unclear - The LLM name, version, and date of inquiry were not clearly specified or provided partially. No - The LLM name, version, and date of inquiry were not reported at all.
	Prompt of the inquiry (e.g., "Please answer..." or "Please role-play as a doctor and respond...")	Yes - The prompt of the inquiry was reported. Unclear - The prompt of the inquiry was not clearly specified or provided partially. No - The prompt of the inquiry was not reported at all.
Reference standard	Is the reference standard likely to correctly classify the target condition?	Yes - If the reference standard is clearly clarified by histopathology results, or follow up, or expert consensus. Unclear - The reference standard is mentioned, but it is not clearly described, or insufficient information is provided to assess its validity. No - If the reference standard is not clarified.

	Who is the evaluator? Number of the evaluator (e.g., Internal medicine doctor, Pediatrician, Radiologist)	Yes - The study clearly identifies the evaluator(s), including their qualifications (e.g., Internal medicine doctor, Pediatrician, Radiologist) and the number of evaluators involved. Unclear - The study mentions evaluators but does not provide sufficient details about their qualifications, expertise, or the number of evaluators. No - The study does not report who the evaluators are or how many evaluators were involved.
	Is the interpretation of the gold standard result blinded? Is the expert assessor unaware of the origin of the diagnostic results (e.g., human vs. LLM-generated)?	Yes - The interpretation of the gold standard result was blinded, meaning the assessor was unaware of whether the diagnostic result was from human or LLM. Unclear - The study does not specify whether blinding was implemented or provides insufficient information to determine this. No - The interpretation of the gold standard result was not blinded, and the assessor had knowledge of the source of the diagnostic results.
Flow and timing	Was the timing of case acquisition and LLM access clearly reported?	Yes - The study clearly reported both the timeframe for case acquisition and the timing of LLM access, ensuring the timeline of events is well-documented. Unclear - The study does not provide sufficient information to determine the timing of case acquisition and/or LLM access. No - The study did not specify the timing for either case acquisition or LLM access.
	Did all cases receive a reference standard?	Yes - All patients received the same gold standard. Unclear - If not clearly reported. No - Patients received more than one or different gold standards.
	Were all cases included in the analysis?	Yes - All cases assessed by both human and LLM were included in the analysis. Unclear - This study did not mention the number of cases included. No - Not all cases assessed by both human and LLM were included in the analysis.
Concerns regarding applicability		

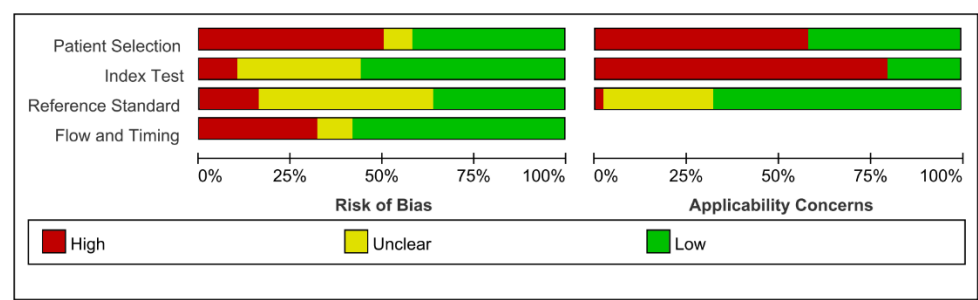
Participant selection	Is there concern that the included patients do not match the review question?	High - The cases are not from a consecutive or random sample of cases or inappropriate exclusions. Unclear - If no details are provided. Low - The cases are from a consecutive or random sample of cases, with appropriate exclusions when necessary.
Index tests	Is there concern that the LLM input scenario in this study differs from real clinical settings?	High - The input to the LLM differs from real clinical settings. Unclear - The study does not provide sufficient information about the style of the inputs provided to the LLM. Low - The input to the LLM closely resembled the real clinical settings.
Reference standard	Is there concern that the gold standard does not appropriately classify the target condition, or the evaluators are not qualified to ensure its reliability?	High - The gold standard is not clearly defined, and/or the study does not report the qualifications, expertise, or number of evaluators involved, raising concerns about the reliability of the classification. Unclear - If details of the gold standard or evaluators are not sufficiently provided. Low - If the reference standard is clearly clarified by histopathology results, or follow up, or expert consensus, and the evaluators are clearly clarified with qualifications and numbers.

Figure S1: Diagnostic accuracy of HCPs with assistance from different LLMs versus HCPs without assistance



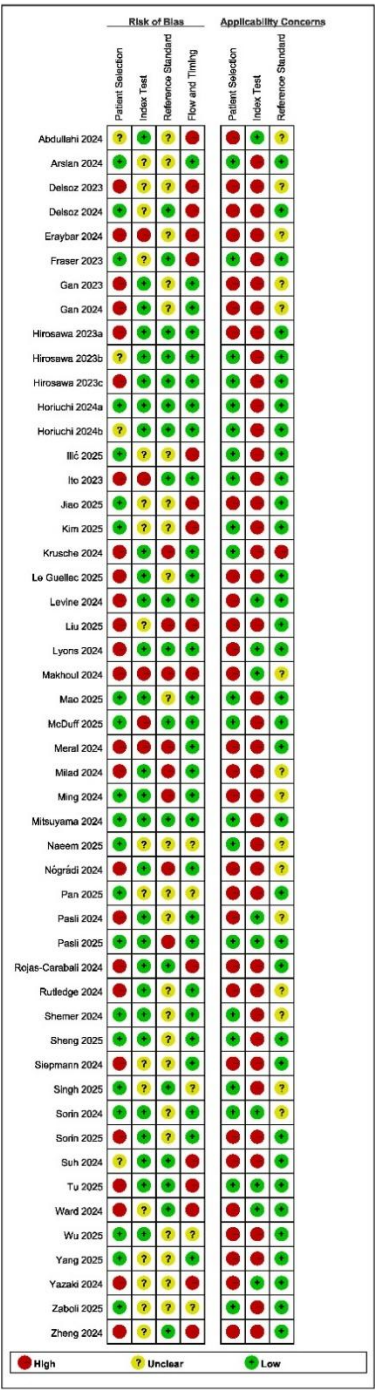
Relative accuracy represents the performance of HCP under different LLM assistance compared with unassisted performance. Values are sorted in descending order of relative accuracy. Comparisons are shown for: a) top-1 diagnosis; b) top-3 diagnosis; c) top-5 diagnosis; d) top-10 diagnosis. LLM, large language model; HCP, healthcare professional. This figure was generated using R (version 4.5.1).

Figure S2: Modified QUADAS-2 summary plot



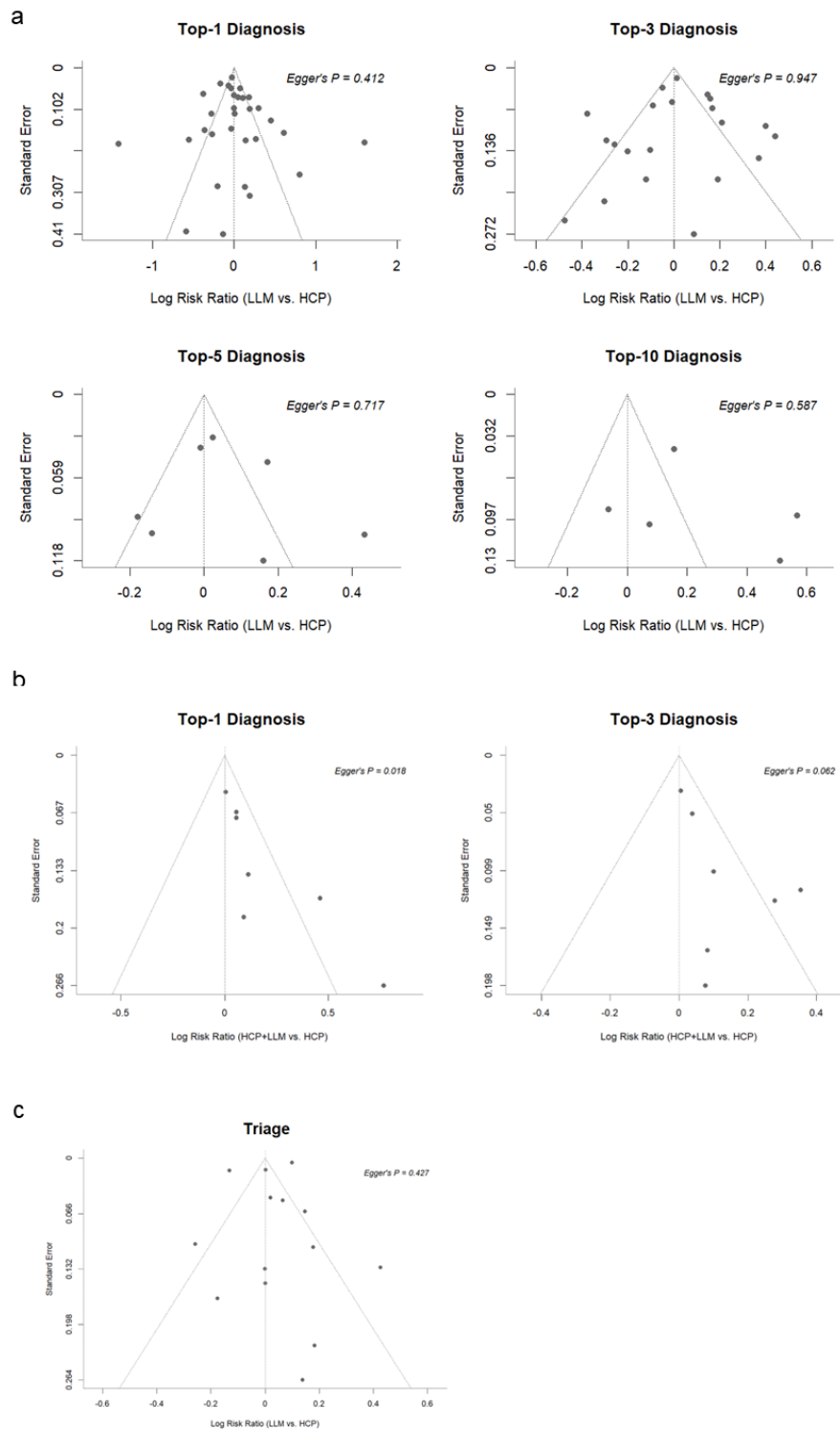
The summary plot displays the distribution of studies classified as high, unclear, or low risk of bias or applicability concerns. Red columns indicate studies with high risk of bias or applicability concerns, yellow columns indicate those with unclear risk of bias or applicability concerns, and green columns indicate studies with low risk of bias or applicability concerns. This figure was generated using RevMan (version 5.3).

Figure S3: Risk of bias and concern of applicability for each item in included studies



The risks of bias for all included studies were evaluated across four domains: patient selection, index test, reference standard, and flow and timing. Applicability concerns were assessed in the following three domains: patient selection, index test, and reference standard. A standardized visual coding system was applied, with red circles indicating high risk of bias or applicability concerns, yellow circles indicating unclear risk of bias or applicability concerns, and green circles indicating low risk of bias or applicability concerns. This figure was generated using RevMan (version 5.3).

Figure S4: Publication bias



Egger test was used to assess publication bias. a. Relative diagnostic accuracy of LLM compared to HCP; b. Relative diagnostic accuracy of HCP with and without LLM assistance; c. Relative triage accuracy of LLM compared to HCP. This figure was generated using R (version 4.5.1).

Table S1: Subgroup analysis of diagnostic and triage accuracy for LLMs and HCPs

	No. of study	LLM (95% CI)	P_{LLM}	HCP (95% CI)	P_{HCP}	RR (95% CI)	P_{RR}
HCP experience							
Top 1 diagnosis			NA		< 0.001		0.030
Junior	9	60.91 (39.94, 78.50)		62.61 (49.48, 74.10)		0.95 (0.80, 1.14)	
Senior				75.94 (65.06, 84.26)		0.77 (0.63, 0.94)	
Top 3 diagnosis			NA		0.002		0.345
Junior	7	67.72 (46.02, 83.77)		73.48 (58.99, 84.22)		0.85 (0.65, 1.10)	
Senior				81.47 (71.48, 88.52)		0.76 (0.59, 0.97)	
Triage			NA		0.123		0.117
Junior	3	71.00 (39.44, 90.20)		70.51 (23.93, 94.78)		1.12 (0.74, 1.68)	
Senior				74.87 (29.41, 95.52)		1.00 (0.79, 1.28)	
Case urgency							
Top 1 diagnosis			0.852		0.046		0.056
Non-urgent	3	83.62 (48.72, 96.48)		88.83 (50.25, 98.43)		1.04 (0.92, 1.17)	
Urgent		71.13 (47.00, 87.25)		80.74 (73.02, 86.65)		0.88 (0.64, 1.22)	
Top 3 diagnosis			0.135		0.379		0.223
Non-urgent	2	91.47 (76.20, 97.29)		97.23 (82.77, 99.61)		0.97 (0.85, 1.11)	
Urgent		80.04 (70.15, 87.24)		92.79 (82.30, 97.27)		0.88 (0.74, 1.04)	
Triage			0.257		< 0.001		0.103
Non-urgent	9	72.90 (55.47, 85.32)		87.84 (73.65, 94.91)		0.88 (0.77, 1.00)	
Urgent		83.17 (54.10, 95.39)		66.03 (52.44, 77.41)		1.04 (0.72, 1.49)	
Case modality							
Top 1 diagnosis			NA		NA		NA
Image	1	47.50 (32.02, 62.98)		60.00 (44.82, 75.18)		0.79 (0.52, 1.20)	

Image and text		67.50 (52.99, 82.01)		70.00 (55.80, 84.20)		0.96 (0.72, 1.30)	
Top 3 diagnosis			NA		NA		NA
Text		39.76 (30.85, 49.41)		16.98 (9.08, 29.53)		2.41 (1.26, 4.59)	
Image	1	5.97 (2.70, 12.68)		41.51 (29.12, 55.07)		0.20 (0.08, 0.55)	
Image and text		35.87 (27.32, 45.42)		49.06 (35.98, 62.26)		0.74 (0.48, 1.14)	
Examination result availability							
Top 3 diagnosis			NA		NA		NA
No	1	53.97 (41.66, 66.28)		73.02 (62.06, 83.98)		0.74 (0.56, 0.97)	
Yes		68.25 (56.76, 79.75)		88.89 (81.13, 96.65)		0.77 (0.64, 0.93)	

All analyses are based on within-study comparisons. Results are presented only for outcomes where primary studies reported sufficient data for synthesis. Consequently, not all metrics (top-1, top-3, top-5, top-10, and triage accuracy) are included in every subgroup category. P_{LLM} tests for significant differences in LLM accuracy across subgroups. P_{HCP} tests for significant differences in HCP accuracy across subgroups. P_{RR} tests whether the relative accuracy of LLMs versus HCPs significantly differs between the subgroups. For the "case modality" and "examination result availability" analyses, P-values are marked as NA when data for a subgroup comparison were available from only a single study, as a formal meta-analytic test could not be performed. RR, Relative accuracy of LLM versus HCP; LLM, large language model; HCP, healthcare professional; NA, Not Applicable.

Table S2: Diagnostic accuracy of HCPs with and without LLM assistance stratified by study design

	No. of study	HCP with LLM (95% CI)	$P_{\text{HCP w LLM}}$	HCP (95% CI)	P_{HCP}	RR (95% CI)	P_{RR}
Top 1 diagnosis			0.069		0.008		0.004
Retrospective clinical cases	4	69.29 (52.27, 82.30)		66.23 (47.63, 80.88)		1.03 (0.97, 1.09)	
Published case reports	3	44.09 (24.60, 65.59)		28.37 (13.72, 49.66)		1.49 (1.04, 2.12)	
Top 3 diagnosis			0.199		0.206		0.132
Retrospective clinical cases	4	74.15 (45.99, 90.62)		68.61 (41.01, 87.29)		1.03 (0.97, 1.09)	
Published case reports	3	53.26 (36.00, 69.77)		45.11 (26.17, 65.58)		1.21 (1.00, 1.46)	
Top 10 diagnosis			0.957		0.020		0.063
Retrospective clinical cases	1	52.00 (40.78, 63.02)		48.00 (36.98, 59.22)		1.08 (0.79, 1.49)	
Published case reports	1	51.66 (46.02, 57.25)		33.44 (28.35, 38.96)		1.54 (1.27, 1.87)	

$P_{\text{HCP w LLM}}$ tests for significant differences in accuracy of HCP with LLM assistance across study design. P_{HCP} tests for significant differences in HCP accuracy across study design. P_{RR} tests whether the relative accuracy of HCP with versus without LLM assistance significantly differs across study design. Top-5 diagnosis is not included because only one study reported this outcome, preventing subgroup analysis. RR, Relative accuracy of HCP with versus without LLM assistance; LLM, large language model; HCP, healthcare professional; NA, Not Applicable.

Table S3: Sensitivity analysis of diagnostic and triage accuracy of LLMs versus HCPs stratified by study risk of bias

	Top 1 diagnosis			Top 3 diagnosis			Top 5 diagnosis			Top 10 diagnosis			Triage		
	n	RR (95% CI)	<i>P</i>	n	RR (95% CI)	<i>P</i>	n	RR (95% CI)	<i>P</i>	n	RR (95% CI)	<i>P</i>	n	RR (95% CI)	<i>P</i>
Participant selection			0.214			0.703			0.325			0.019			0.554
Low	13	0.94 (0.78, 1.13)		11	0.91 (0.79, 1.06)		3	1.14 (0.88, 1.49)		2	1.53 (1.17, 2.00)		4	0.98 (0.87, 1.11)	
High/unclear	19	0.84 (0.72, 0.99)		10	0.93 (0.82, 1.06)		4	0.95 (0.77, 1.19)		3	1.00 (0.72, 1.38)		10	1.03 (0.93, 1.14)	
Index tests			0.059			0.370			0.325			0.126			0.990
Low	20	0.80 (0.69, 0.94)		16	0.88 (0.78, 0.99)		4	0.95 (0.77, 1.19)		4	1.06 (0.80, 1.40)		6	1.01 (0.89, 1.15)	
High/unclear	12	1.03 (0.86, 1.23)		5	1.02 (0.83, 1.25)		3	1.14 (0.88, 1.49)		1	1.76 (1.46, 2.12)		8	1.01 (0.91, 1.13)	
Reference standard			0.107			0.449			0.929			0.583			0.278
Low	17	0.81 (0.69, 0.96)		10	0.85 (0.72, 1.01)		6	1.04 (0.85, 1.26)		4	1.16 (0.79, 1.68)		5	0.96 (0.86, 1.08)	
High/unclear	15	0.98 (0.83, 1.15)		11	0.95 (0.86, 1.05)		1	1.02 (0.97, 1.09)		1	1.39 (0.97, 1.99)		9	1.05 (0.94, 1.17)	
Flow and timing			0.932			0.722			0.838			0.992			0.581
Low	18	0.88 (0.72, 1.06)		14	0.91 (0.78, 1.06)		4	1.01 (0.73, 1.38)		4	1.18 (0.80, 1.72)		9	1.03 (0.95, 1.12)	
High/unclear	14	0.96 (0.80, 1.15)		7	0.93 (0.82, 1.05)		3	1.06 (0.96, 1.17)		1	1.17 (1.08, 1.27)		5	0.99 (0.83, 1.19)	

n, number of study; RR, Relative accuracy of LLM compared to HCP; *P* tests for significant differences in relative accuracy across study risk of bias.

Table S4: Sensitivity analysis of diagnostic accuracy of HCPs with and without LLM assistance stratified by study risk of bias

	Top-1 diagnosis			Top-3 diagnosis		
	n	RR (95% CI)	<i>P</i>	n	RR (95% CI)	<i>P</i>
Participant selection			0.364			0.467
Low	5	1.25 (0.99, 1.59)		5	1.15 (1.00, 1.33)	
High/unclear	2	1.06 (0.96, 1.16)		2	1.04 (0.95, 1.15)	
Index tests			0.994			0.222
Low	4	1.17 (0.92, 1.50)		4	1.02 (0.96, 1.07)	
High/unclear	3	1.15 (0.94, 1.41)		3	1.22 (0.99, 1.50)	
Reference standard			0.050			0.104
Low	2	1.32 (0.94, 1.85)		2	1.25 (0.97, 1.60)	
High/unclear	5	1.04 (0.98, 1.11)		5	1.03 (0.97, 1.09)	
Flow and timing			0.517			0.151
Low	5	1.09 (0.97, 1.22)		6	1.08 (0.98, 1.19)	
High/unclear	2	1.44 (0.72, 2.88)		1	1.32 (1.03, 1.69)	

n, number of study; RR, Relative accuracy of HCP with versus without LLM assistance; *P* tests for significant differences in relative accuracy across study risk of bias. Sensitivity analyses for Top-5 (n=1) and Top-10 (n=2) outcomes were omitted as the number of studies was insufficient for subgroup analysis.