

Supplementary Appendix

Diagnostic accuracy of large language models for rare diseases: a systematic review and meta-analysis

Minh Ha Nguyen, Chih-Ting Yang, Thomas A. Cassini, Fan Ma, Rizwan Hamid, Lisa Bastarache, Josh F. Peterson, Hua Xu, Lingyao Li, Siyuan Ma, and Cathy Shyr

Correspondence to: Cathy Shyr, Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN 37203, U.S.A. Email: cathy.shyr@vumc.org

Web appendix to:

Diagnostic accuracy of large language models for rare diseases: a systematic review and meta-analysis

Contents

1	Search strategy	1
1.1	PubMed/MEDLINE	1
1.2	arXiv and medRxiv	1
2	Risk of bias assessment	2
3	Supplementary methods	3
3.1	Phenopacket Store causal variant verification	3
3.2	Orphanet prevalence mapping methodology	3
4	Supplementary figures	4
5	Supplementary tables	6
6	PRISMA-DTA checklist	17

Search strategy

The search strategy combined two concept blocks across all databases. Block 1 (LLM terms) and Block 2 (rare disease terms) were combined with AND. The PubMed/MEDLINE strategy is shown below; equivalent strategies were adapted for Embase, Web of Science Core Collection, and Cochrane Library using database-specific syntax and controlled vocabulary where applicable. All searches covered January 2020 to February 2026 with no language restrictions.

PubMed/MEDLINE

Block 1 — LLM and AI terms:

("large language model"[tiab] OR "LLM"[tiab] OR "LLMs"[tiab] OR "GPT-4"[tiab] OR "GPT-3.5"[tiab] OR "ChatGPT"[tiab] OR "Gemini"[tiab] OR "LLaMA"[tiab] OR "PaLM"[tiab] OR "Med-PaLM"[tiab] OR "DeepSeek"[tiab] OR "Qwen"[tiab] OR "generative AI"[tiab] OR "foundation model"[tiab] OR "language model"[tiab] OR "retrieval-augmented generation"[tiab] OR "agentic"[tiab] OR "AI agent"[tiab] OR "multi-agent"[tiab] OR "Natural Language Processing"[MeSH])

Block 2 — Rare disease terms:

("rare disease"[tiab] OR "rare diseases"[tiab] OR "rare genetic disease"[tiab] OR "orphan disease"[tiab] OR "Mendelian disease"[tiab] OR "Mendelian disorder"[tiab] OR "monogenic disease"[tiab] OR "undiagnosed disease"[tiab] OR "undiagnosed patient"[tiab] OR "genetic syndrome"[tiab] OR "inborn errors"[tiab] OR "diagnostic odyssey"[tiab])

A supplementary MeSH sensitivity search added "Rare Diseases"[MeSH] to Block 2. This was applied to PubMed only; the Emtree equivalent inflated Embase results 4-6-fold, confirming that controlled vocabulary addition would compromise specificity in other databases.

arXiv and medRxiv

For preprint servers (arXiv sections: cs.CL, cs.AI, cs.LG, q-bio; medRxiv via Crossref API), nine keyword queries were used: "rare disease" AND "LLM"; "rare disease" AND "large language model"; "rare disease diagnosis" AND "agent"; "genetic diagnosis" AND "language model"; "phenotype" AND "LLM" AND "diagnosis"; "clinical genetics" AND "GPT"; "differential diagnosis" AND "rare" AND "AI"; "variant interpretation" AND "LLM"; "HPO" AND "LLM". Results were deduplicated across queries and manually reviewed.

Risk of bias assessment

Risk of bias was assessed using a modified QUADAS-3 instrument adapted for LLM-based diagnostic evaluation studies. The standard QUADAS-3 framework comprises four domains assessing participant selection, index test conduct, target condition definition, and flow and timing. We supplemented these with three AI-specific domains to address methodological concerns unique to LLM evaluations: data leakage, reproducibility, and evaluation fairness.

Domain 1 — Participant selection assessed whether the evaluation cohort was representative of the intended clinical population or whether case selection introduced spectrum bias (e.g., restricting to well-characterised diseases or enriching for diseases well-represented in LLM training data).

Domain 2 — Index test conduct evaluated whether the LLM system was applied as intended, including whether the prompt, model version, and configuration were consistent across all evaluation cases and whether the system had access to the reference standard diagnosis during evaluation.

Domain 3 — Target condition definition assessed whether the ground-truth diagnosis was established using an appropriate reference standard (e.g., molecular genetic confirmation) applied independently of the LLM output.

Domain 4 — Flow and timing evaluated whether all enrolled cases were included in the analysis and whether there was appropriate handling of inconclusive or failed LLM outputs.

Domain 5 — Data leakage (AI-specific) assessed the risk that evaluation cases or their diagnostic labels were present in the LLM's pretraining data, retrieval knowledge base, or fine-tuning dataset. This domain was rated high risk when evaluation benchmarks predated the LLM's training cutoff without temporal holdout validation, or when the system's knowledge base contained the benchmark's source data.

Domain 6 — Reproducibility (AI-specific) assessed whether sufficient methodological detail was provided to reproduce the evaluation, including exact model version, temperature and sampling parameters, prompt text, and code and data availability.

Domain 7 — Evaluation fairness (AI-specific) assessed whether all compared systems received equivalent prompting effort and whether the evaluation metric (R@1) was applied identically across systems, including consistency in disease name matching.

Each domain was rated as low risk (–), high risk (+), or unclear (?). An overall judgment of high risk of bias was assigned if any domain was rated high. Assessments were performed independently by two reviewers, with disagreements resolved by discussion.

Supplementary methods

Phenopacket Store causal variant verification

To characterise the information embedded in Phenopacket Store cases used by three included studies (Reese 2024, Chimirri 2025, MINT 2025), we programmatically inspected all phenopacket files in the repository (monarch-initiative/phenopacket-store, v0.1.26, GA4GH Phenopacket Schema v2.0.2). A total of 9,588 JSON files were inspected. Every case contained at least one pre-identified causal variant (mean 1.1 per case; 90.0% with one variant, 10.0% with two variants consistent with compound heterozygous inheritance). Of 10,547 variant entries, 94.4% included full VCF-format records (chromosome, position, reference and alternate alleles) and 99.9% carried an ACMG classification of pathogenic or likely pathogenic. This confirms that the Phenopacket Store embeds molecular diagnostic information alongside phenotypic descriptions, which is relevant when interpreting performance differences between studies that used HPO-only input versus those that incorporated the full phenopacket content.

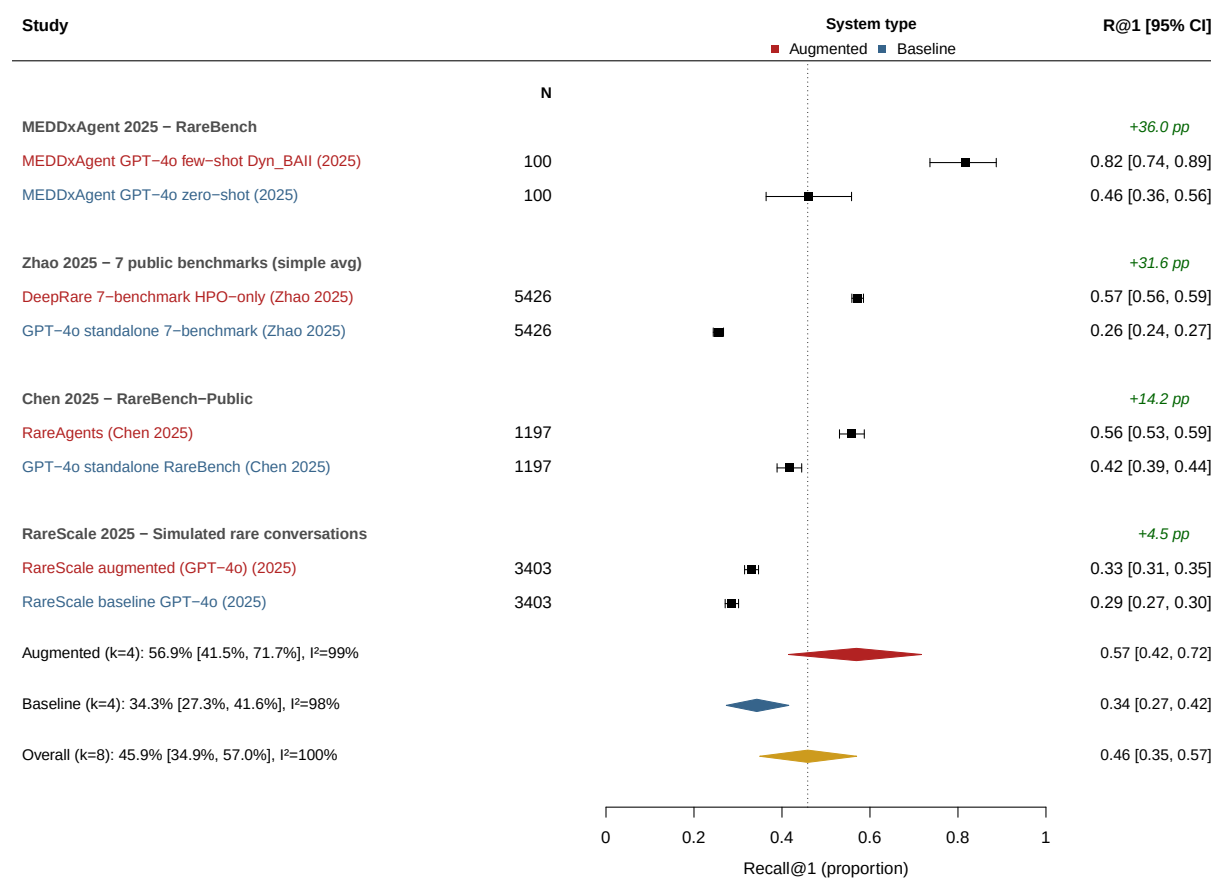
Orphanet prevalence mapping methodology

To investigate whether benchmark disease composition contributed to observed performance heterogeneity, we mapped diseases from four commonly used evaluation benchmarks to Orphanet prevalence classifications using Orphanet Product 9 (epidemiology data, accessed March 2026).

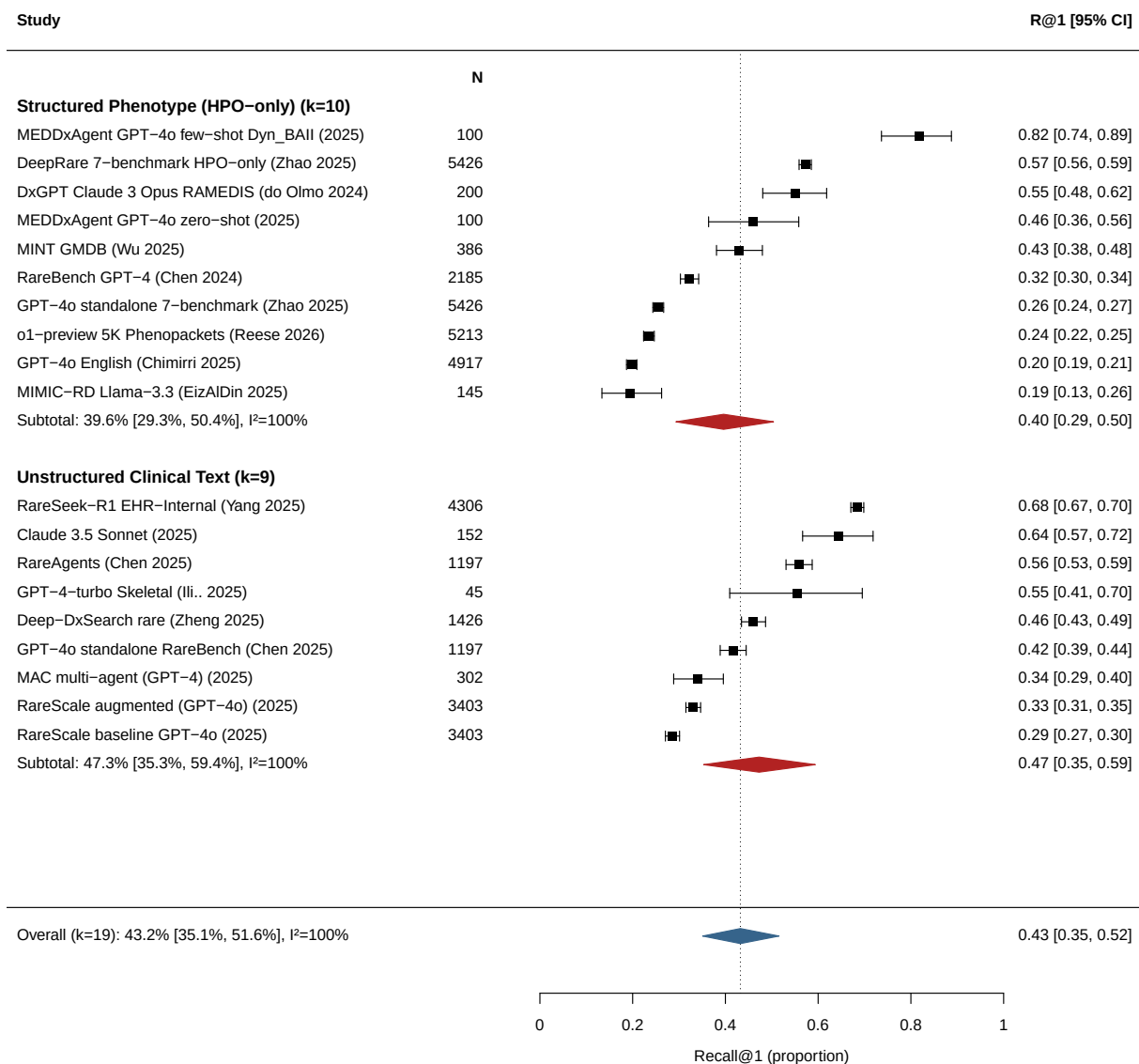
For the Phenopacket Store (694 unique diseases), OMIM disease identifiers were extracted from 9,588 GA4GH phenopacket JSON files and cross-referenced to OrphaCode identifiers via the OMIM-to-OrphaCode mapping table (mapping rate: 49.1%). For RareBench (301 diseases), MAC Medline (301 diseases), and the Chinese Rare Disease List (66 diseases), disease names were mapped to Orphanet entries via programmatic search and manual curation (mapping rates: 88–91%).

Mapped diseases were classified according to Orphanet point prevalence categories: ultra-rare (<1 per 1,000,000), rare (1–9 per 1,000,000), or higher-prevalence (1 per 100,000). Diseases with multiple prevalence estimates were assigned their most specific geographic match. The resulting prevalence distributions showed substantial differences across benchmarks: the Phenopacket Store comprised 52.8% ultra-rare diseases, compared with 29.3% for RareBench, 30.1% for MAC Medline, and 1.7% for the Chinese Rare Disease List.

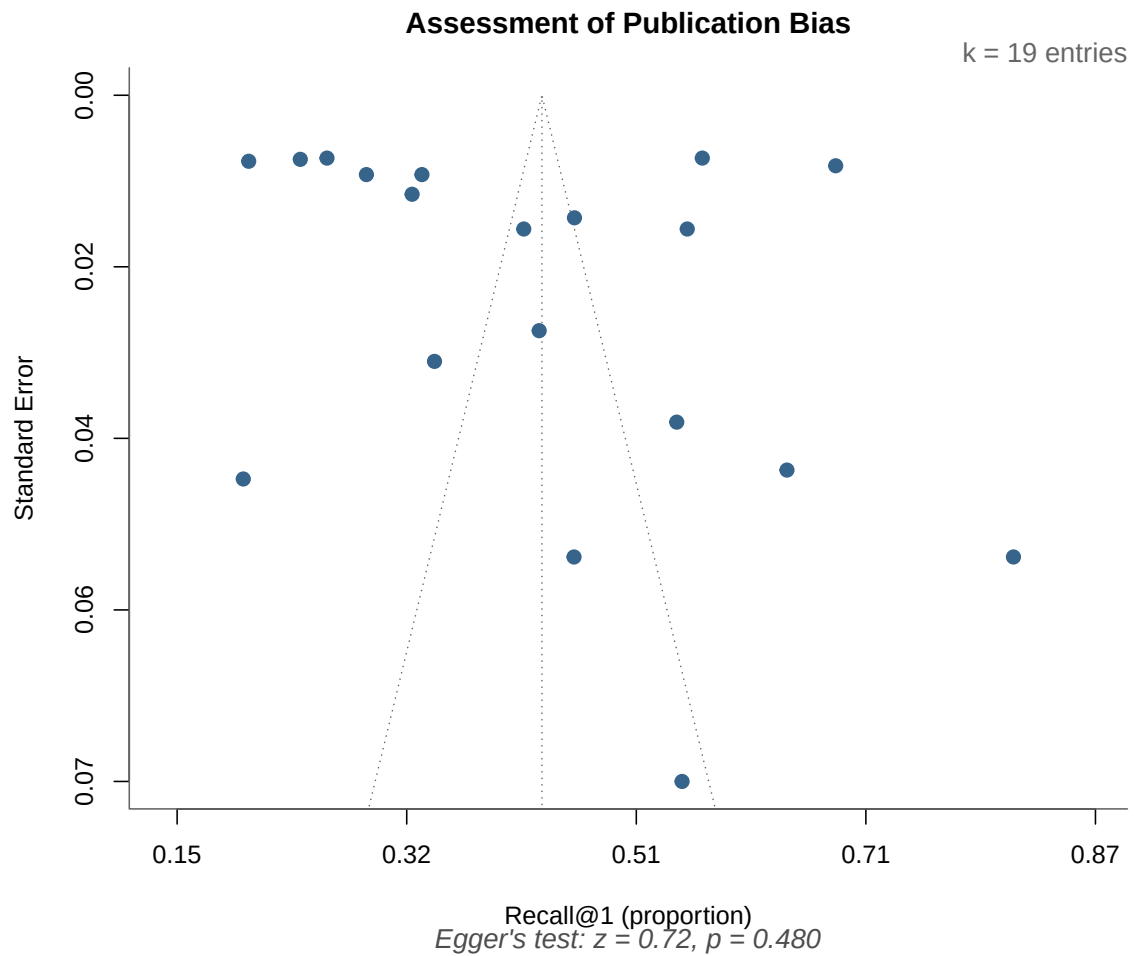
Supplementary figures



Supplementary Figure S1: Forest plot of paired within-study comparisons of augmented versus baseline LLM systems. Four studies provided matched comparisons on the same benchmark. In all four studies, augmented systems achieved higher R@1 than their baseline counterparts, with improvements ranging from 4.5 to 36 percentage points.



Supplementary Figure S2: Forest plot of subgroup analysis by input modality (HPO-only versus clinical text).



Supplementary Figure S3: Funnel plot for assessment of publication bias and small-study effects. The plot shows the distribution of study effect sizes against standard errors. Visual inspection does not suggest appreciable asymmetry.

Supplementary tables

Of the 15 included studies, 8 (53%) reported some form of prompt text. Three studies (20%) published complete verbatim prompt sets: DeepRare (14 numbered prompts in Supplementary Information Section 1.8), DxGPT (3 prompts in Appendix), and RareAgents (41 specialist system messages in code repository). Among the 8 augmented systems, 5 used explicit clinical role assignment, and 7 incorporated chain-of-thought or iterative reasoning. External knowledge was injected into prompts by 8 studies (53%). Two studies (MEDDxAgent, RareScale) provided explicit candidate disease lists. Four studies (DeepRare, MEDDxAgent, RareBench, Chimirri) included few-shot examples. Two studies (RareSeek, MINT) used fine-tuning rather than prompt engineering.

Supplementary Table S1. Prompting strategies and prompt component taxonomy for 15 included studies. Taxonomy columns: Role = clinical expert persona assigned; CoT = chain-of-thought or iterative reasoning; Out = output format specified; Cand = candidate disease list provided; Kn = external knowledge injected; Ex = few-shot examples. Y = present, N = absent. † = encoded implicitly via fine-tuning. ‡ = explicitly stated as absent.

Study	Prompt	Role	CoT	Out	Cand	Kn	Ex
Zhao (DeepRare)	<i>14 prompts published verbatim in Nature Supplementary Information Section 1.8 (pp. 7–17). Key prompts:</i>						
Prompt 1: Baseline diagnosis	“You are a specialist in the field of rare diseases. You will be provided and asked about a complicated clinical case; read it carefully and then provide a diverse and comprehensive differential diagnosis. Patient’s {info_type}: {patient_info}. Enumerate the top 5 most likely diagnoses. Be precise, listing one diagnosis per line.”	Y	N	Y	N	N	N
Prompt 2: Evaluation	“You are a specialist in the field of rare diseases. I will now give you five predicted diseases. Please identify the rank of the following gold-standard diagnosis. Please output the predicted rank; otherwise, output ‘No’. Only output ‘No’ or ‘1-5’ numbers.”	Y	N	Y	N	N	N
Prompt 3: Rare/common discriminator	“You are a specialist in the field of rare diseases. Given the following HPO phenotypes: {patient_info}. Is this most likely associated with a rare disease or a common disease? Please reply with only one word: ‘rare’ or ‘common’.”	Y	N	Y	N	N	N
Prompt 4: Disease type classification	“You are a medical disease classifier specialising in categorising diseases into predefined categories.” Classifies diseases into 14 body system categories (Blood/Heart/Circulation, Bones/Joints/Muscles, Brain/Nerves, etc.). Output: JSON format only with disease name and category array.	Y	N	Y	N	N	N
Prompt 5: Tentative diagnosis	“You are a specialist in the field of rare diseases. You have access to the following context: Online knowledge: {web_diagnosis}; LLM-generated diagnoses: {llm_response}; Diagnosis API results: {api_response}; Similar cases: {similar_cases}. Enumerate the top 5 most likely rare disease diagnoses with diagnostic reasoning and in-text citations.”	Y	Y	Y	N	Y	N
Prompt 6: Gene analysis	“You are a specialist in the field of rare diseases. Here is a rare disease diagnosis case. Exomiser summary: {exomiser_summary}; HPO terms: {hpo_terms}; Preliminary diagnosis: {pheno_diagnosis}. Enumerate the top 5 most likely rare disease diagnoses, prioritising the Exomiser gene/variant results.”	Y	Y	Y	N	Y	N
Prompt 7: Disease reflection	“Assume you are a doctor specialised in rare disease diagnosis. Evaluate whether the proposed diagnosis is correct for this patient. Begin with ‘DIAGNOSIS ASSESSMENT: [Correct/Incorrect]’. Structure: 1. Patient summary; 2. Proposed diagnosis analysis; 3. References. Patient phenotype: {patient_info}. Medical literature: {disease_knowledge}.”	Y	Y	Y	N	Y	N
Prompt 8: Final diagnosis	“Patient presentation: {patient_info}. Similar cases: {similar_cases}. Primary diagnosis results: {tentative_result}. Disease Reflection: {judgements}. Task: Enumerate the top 5 most likely rare disease diagnoses with 3–4 sentences of diagnostic reasoning and in-text citations.”	N	Y	Y	N	Y	N
Prompt 9: HPO extraction	“Given a paragraph of patient information from discharge note, please extract the phenotype about this patient only. Check the HPO database. Only output the extracted phenotypes. Use the format: {‘HPO’: ‘HP:0000000’, ‘Phenotype’: ‘description’}.”	N	N	Y	N	N	N
Prompt 10: HPO normalisation	“You are a medical terminology translator specialising in rare diseases. Convert patient phenotype description into standardised HPO concept. Output JSON format only.” (Includes 2 worked examples)	Y	N	Y	N	N	Y
Prompt 11: Knowledge summarisation	“Assume you are a doctor, please summarise these medical articles into a paragraph, only keep key message, mainly focus on the phenotype and related disease.”	Y	N	Y	N	N	N
Prompt 12: Case comparison	“Assume you are a doctor experienced in rare disease diagnosis. Please judge if the two patient cases are likely to be the same disease based on the patient information. Only output ‘Yes’ or ‘No’. Patient 1: {patient_info}. Patient 2: {retrieved_case}.”	Y	N	Y	N	N	N

Table S1 continued

Study	Prompt	Role	CoT	Out	Cand	Kn	Ex
Prompt 13: Zero-shot LLM inference	“You are a specialist in the field of rare diseases. You will be provided and asked about a complicated clinical case; read it carefully and then provide a diverse and comprehensive differential diagnosis. Patient’s {info_type}: {patient_info}. Enumerate the top 5 most likely diagnoses. Each diagnosis should be a rare disease. Use ** to tag the disease name.”	Y	N	Y	N	N	N
Prompt 14: Rare disease judgment	“You are a specialist in the field of rare diseases. Given a disease name, determine whether it is a rare disease or a common disease. Output can only be 1/0.” (Includes 4 worked examples)	Y	N	Y	N	N	Y
Chen X (RareAgents), 2024	<i>41 specialist system messages published in code repository. Example:</i>						
MDT specialist	“As a [Specialist], you are dedicated to [specialist-specific clinical description]. Your expertise involves diagnosing and treating [domain]. Demonstrate proficiency in [domain assessments]. Communicate effectively with patients and families.” (41 predefined departments)	Y	N	N	N	N	N
Diagnostic reasoning	Multi-round MDT discussion (up to R rounds). Phenomizer p-values, LIRICAL posteriors, and Phenobrain scores injected as tool outputs. Dynamic memory retrieves top-k=5 similar cases.	Y	Y	Y	N	Y	N
Chen X (MAC), 2025	Admin presents clinical case. 2–5 doctor agents provide diagnostic reasoning. Supervisor moderates up to 13 conversation rounds until consensus. Prompt text not published; AutoGen framework used.	Y	Y	Y	N	N	N
Rose (MEDDx-Agent), 2025	CoT prompting with dynamic few-shot examples selected via BGE-BASE-EN-v1.5 embeddings. “You may be provided a list of diagnosis options you can choose from. You must use this exact disease terminology.” 102 unique diseases from RareBench subset. Wikipedia + PubMed knowledge retrieval.	N	Y	Y	Y	Y	Y
Zheng (Deep-DxSearch), 2025	System: “You are an AI assistant specialised in diagnosing rare diseases based on phenotypes. Your task is to analyse patient phenotypes, match them with similar cases, and identify the most likely diagnosis.” Tools: Phenotype Matching (<match>), Disease Knowledge Search (<search>). Workflow: submit phenotypes → analyse matches → reason in <think> → query diseases → synthesise in <answer>.	Y	Y	Y	N	Y	N
Yang (RareSeek), 2025	3-stage fine-tuning pipeline: (1) domain-specific instruction tuning, (2) CoT fine-tuning on 17,477 reasoning chains, (3) GraphRAG integration with Neo4j knowledge graph (635K ClinVar/HGMD variants, 7.5K OMIM entries, 4.3K Orphanet links, 16.7K HPO phenotypes). Generates up to 20 candidates normalised to Orphanet via MONDO API.	Y†	Y†	Y	N	Y†	N
Wang (MINT), 2026	Llama-3.2-3B-Instruct aligned via ORPO. Preference pairs from multimodal teacher (GestaltMML): “chosen” = teacher’s top-ranked disease, “rejected” = lower-ranked. No explicit prompt engineering; 6.5K GMDb training samples. Input: HPO terms + demographics.	N†	N	Y	N	Y†	N
Schumacher (RareScale), 2025	<i>Three-stage prompt structure described but not published verbatim:</i>						
Stage 1: Baseline DDx	“Generate 5 diagnoses without considering the rare candidate list.”	N	N	Y	N	N	N
Stage 2: Candidate-augmented	“Consider the candidate list (but not required to use them; should discard if not appropriate).” Specialised 7B model provides 5-disease candidate list.	N	Y	Y	Y	Y	N
Stage 3: Final selection	“Select 5 most appropriate from both sets, including discarding all rare candidates if best.”	N	Y	Y	Y	Y	N
Chen X (RareBench), 2024							
Zero-shot baseline	System: “You are a specialist in the field of rare diseases.” Hit@k (k=1,3,10). CoT explored (32.3% zero-shot → 33.2% CoT).	Y	Y	Y	N	N	N
Dynamic few-shot	IC-weighted knowledge graph embeddings (HPO, OMIM, Orphanet; 17.2K phenotypes, 9.26K diseases) select m most similar training examples. m=3 boosted smaller models’ R@1 by 108%.	Y	Y	Y	N	Y	Y
Aldin et al., 2026	“Given a patient’s list of observed phenotypes, LLMs are asked to predict a list of the 10 most likely rare diseases in order of likelihood.” Prompt text not published.	N	N	Y	N	N	N

Table S1 continued

Study	Prompt	Role	CoT	Out	Cand	Kn	Ex
Chimirri et al., 2025	phenopacket2prompt generates clinical vignette from GA4GH Phenopacket JSON (age, sex, observed/excluded HPO terms, onset). “We request the LLMs to return an ordered list of candidate diagnoses, giving an example output.”	N	N	Y	N	N	Y
Ilić et al., 2025	Standardised clinical vignettes (demographics, prenatal history, skeletal findings, radiology, clinical course, biochemistry, family history). “Generate primary diagnosis + 2 differential diagnoses + confidence score (1–5).” Paper states: “No role-playing cues or guiding frameworks were supplied.”	N‡	N	Y	N	N	N
Ao et al., 2025	Clinical data excluding pathognomonic indicators (no genetic tests, biopsies, pathology). “Each model generated the top five most likely diagnoses, ranking them by probability.” Prompt text not published (Letter).	N	N	Y	N	N	N
Reese et al., 2026	phenopacket2prompt converts Phenopacket JSON to clinical narrative (sex, age of onset, age at exam, observed/excluded HPO terms). “Return a differential diagnosis as a list of disease names.” Zero-shot only. Prompts deposited on Zenodo.	N	N	Y	N	N	N
Olmo (DxGPT), 2024	“Behave like a hypothetical doctor who has to do a diagnosis for a patient. Give me a list of potential diseases with a short description. You have to indicate which symptoms the patient has in common with the proposed disease and which symptoms the patient does not have in common. Symptoms: {HPO_terms}”	Y	N	Y	N	N	N

Supplementary Table S2. Per-architecture descriptive subgroup estimates (DerSimonian–Laird random-effects model).

Subgroup	k	N	Pooled R@1	95% CI	95% PI	I ²
Agentic	4	8,351	48.7%	40.9%–56.4%	32.0%–65.5%	97.2%
Knowledge-augmented	2	3,503	58.3%	12.5%–96.4%	0.0%–100.0%	99.0%
LoRA fine-tuned	2	4,692	56.0%	31.2%–79.4%	15.8%–92.0%	98.9%
Standalone	11	22,983	35.3%	30.5%–40.4%	20.2%–52.1%	98.1%
Augmented combined (primary)	8	16,546	52.5%	42.0%–62.9%	22.7%–81.3%	99.4%

Per-architecture pools are reported descriptively. The knowledge-augmented and fine-tuned categories each contain only two entries; their prediction intervals (effectively 0–100% and 15.8–92.0%, respectively) are uninterpretable on their own, which motivates the combined augmented category as the primary contrast in the main text. k = number of system–dataset entries in the pool; N = total number of cases; CI = confidence interval; PI = prediction interval (DerSimonian–Laird random-effects model).

Supplementary Table S3. Row-level data-leakage and model-cutoff evidence for the 19 selected system–dataset entries.

Entry	Model backbone	Benchmark / source	Cutoff vs. source	Leakage tier	D5	Mechanism note
DeepRare 7-benchmark HPO-only (Zhao 2025)	DeepSeek-V3 host; comparators include GPT-4o / Gemini / Claude / o3-mini / DeepSeek-R1 – all versioned	Seven public benchmarks: RareBench subsets / MyGene2 / MIMIC-IV-Rare / DDD / Phenopacket Store	Pre-cutoff public benchmark exposure plus possible tool/retrieval overlap	Likely pretraining exposure	HIGH	Public benchmarks predate or overlap modern frontier-model training corpora; no contamination audit performed.
GPT-4o standalone 7-benchmark (Zhao 2025)	GPT-4o pinned as gpt-4o-2024-11-20; cutoff not extracted in local RoB	Same seven public benchmarks as DeepRare	Pre-cutoff public benchmark exposure	Likely pretraining exposure	HIGH	Standalone baseline; leakage would inflate the baseline and make augmented-vs-standalone gap conservative.
RareAgents (Chen 2025)	Llama-3.1-70B backbone; GPT-4o used as judge and comparator	RareBench-Public; dynamic memory retrieves top-5 similar cases from RareBench itself	Pre-cutoff public benchmark plus retrieval-bank overlaps test set	Direct evidence of inflation	HIGH	Memory ablation cost -22.4 Hit@1; retrieval bank / test-set relationship insufficiently documented.
GPT-4o standalone RareBench (Chen 2025)	GPT-4o standalone; exact API version and cutoff not pinned in local extraction	RareBench-Public	Pre-cutoff public benchmark exposure	Likely pretraining exposure	HIGH	Same benchmark as RareAgents augmented arm; no specific retrieval leak for standalone row.
MAC multi-agent GPT-4 (2025)	GPT-4; version not pinned; temporal design selected post-January 2022 cases	Novel 302-case Medline collection from post-January 2022 reports	Partial temporal mitigation (post-Jan 2022 selection); GPT-4-specific exposure not fully excluded	Partial mitigation	LOW	Best temporal-design example; lowers leakage risk but GPT-4 cutoff vs. collection date leaves residual uncertainty.
MEDDxAgent GPT-4o zero-shot (2025)	GPT-4o; RoB records approximate October 2023 cutoff	RareBench 100-case subset; 102-disease candidate list constraint	Pre-cutoff public benchmark plus constrained answer set	Direct evidence of inflation	HIGH	102-disease restricted answer set inflates absolute R@1 and limits comparability with open-ended systems.
MEDDxAgent GPT-4o few-shot Dyn_BAII (2025)	Same GPT-4o setting as MEDDxAgent zero-shot	Same RareBench subset and same 102-disease candidate list; dynamic few-shot examples added	Pre-cutoff public benchmark plus constrained answer set	Direct evidence of inflation	HIGH	Within-study delta interpretable (both arms share restriction) but absolute 82% R@1 not comparable with open-ended generation.
Deep-DxSearch rare (Zheng 2025)	MedGemma-27B backbone; GPT-4o and DeepSeek-R1 comparators not exactly pinned	In-distribution rare-disease evaluation; retrieval database contains RareBench / MME / LIRICAL / MIMIC-R cases	Retrieval database overlaps with evaluation benchmarks; no explicit query-case exclusion	Direct evidence of inflation	HIGH	Strongest retrieval-overlap row; OOD evaluation more convincing than ID evaluation.
RareSeek-R1 EHR-Internal (Yang 2025)	Fine-tuned DeepSeek-R1-Distill-Llama-70B; baseline LLM versions not fully pinned	Same institution and EHR system as 48852-EHR training corpus; no temporal split described	Same-institution EHR overlap (training ↔ evaluation)	Direct evidence of inflation	HIGH	EHR-Internal R@1 68.4% vs. RareBench 39.2%; 29.2 pp gap – though domain shift and modality are co-confounders.
MINT GMDB (Wu 2025)	Llama 3.2-3B-Instruct with ORPO alignment	GMDB train/test: case-level holdout but disease-level overlap; external Phenopacket validation separates overlapping vs. unseen diseases	Disease-level overlap between training and evaluation distributions	Direct evidence of inflation	HIGH	Overlapping diseases R@1 47.6% vs. disjoint diseases 7.0% – roughly 41 pp gap directly demonstrating memorisation effect.

Table S3 continued

Entry	Model backbone	Benchmark / source	Cutoff vs. source	Leakage tier	D5	Mechanism note
RareScale augmented GPT-4o (2025)	GPT-4o final ranker; smaller candidate generator trained from expert-system/simulated data; GPT-4o version not pinned	Same 575 rare diseases used across generation and evaluation; expert-system-derived answers	Same-disease / no disease-level holdout plus circular expert-system overlap	Direct evidence of inflation	HIGH	Design-overlap example: candidate generator narrows disease space using same expert-system source as evaluation.
RareBench GPT-4 (Chen 2024)	GPT-4 1106-preview; approximate April 2023 cutoff	RareBench Task 4 combines public MME / LIRICAL / HMS / RAMEDIS plus PUMCH; 1122/2185 cases (51%) are public benchmark-derived predating cutoff	Pre-cutoff public benchmark exposure (51% of cases)	Likely pretraining exposure	HIGH	Cleanest cutoff-reasoning example: public components predate model cutoff; no contamination audit performed.
MIMIC-RD Llama-3.3 (EizAlDin 2025)	Llama 3.3 70B; RoB records December 2023 cutoff	MIMIC-IV notes are public/credentialed (2008–2022); MIMIC-RD mapping is novel	Plausible public-dataset exposure partially mitigated by novel mapping and noisy clinical-note input	Likely pretraining exposure	HIGH	Risk is exposure to MIMIC text patterns or derivatives – not published case-report memorisation; phrase cautiously.
GPT-4o English (Chimirri 2025)	gpt-4o-2024-08-06 temperature 1; knowledge cutoff October 2023	Phenopacket Store v0.1.19 cases from 706 publications	Pre-cutoff public benchmark; authors explicitly acknowledge likely training exposure	Likely pretraining exposure	HIGH	Best author-acknowledged example: authors note models likely had access to some source clinical data.
GPT-4-turbo Skeletal (Ilic 2025)	GPT-4-turbo approximate 2024 version; cutoff around April 2024	Specific 45-patient novel confirmed cohort; skeletal dysplasia associations well represented in literature	Novel cohort but domain familiarity (not case-level leakage)	Likely pretraining exposure	HIGH	Lower-specificity concern: domain familiarity rather than exposure to these specific patients.
Claude 3.5 Sonnet Chinese cases (2025)	Claude 3.5 Sonnet; approximate early-2024 cutoff	Chinese Medical Case Repository reports predate cutoff and are public medical literature	Pre-cutoff public case-report exposure	Likely pretraining exposure	HIGH	High R@1 (64.5%) may partly reflect case-report familiarity; exact version and contamination status underdocumented.
o1-preview 5K Phenopackets (Reese 2026)	o1-preview (published EJHG version); exact cutoff not captured in local RoB (local file is stale preprint-era GPT-4)	Phenopacket Store case reports (same benchmark family as Chimirri)	Pre-cutoff public benchmark exposure; exact o1-preview cutoff not captured	Unassessable	HIGH	Retain D5 benchmark rationale; do not cite stale GPT-4 model details – RoB/supplement update needed.
RareScale baseline GPT-4o (2025)	GPT-4o standalone baseline; version not pinned	Same simulated rare conversations and 575-disease expert-system source as RareScale augmented	Same-disease / no disease-level holdout shared with augmented arm	Direct evidence of inflation	HIGH	Baseline shares evaluation source with augmented RareScale; within-paper comparison useful but absolute R@1 has overlap concerns.
DxGPT Claude 3 Opus RAMEDIS (do Olmo 2024)	Claude 3 Opus; exact version and cutoff not pinned in RoB file	RAMEDIS is a long-established public rare-disease benchmark; later part of RareBench	Pre-cutoff public benchmark exposure	Likely pretraining exposure	HIGH	Relative rankings across 13 models more reliable than absolute R@1 – all models share the same prompt/data exposure risk.

Supplementary Table S4. Short rationales for the high-risk domain judgments in the modified QUADAS-3 risk-of-bias assessment of the 19 selected meta-analysis entries.

Study	Selected meta-analysis entry	High-risk domains	Short rationale
Zhao (DeepRare), 2026	DeepRare 7-benchmark HPO-only	D5	Public benchmark estimate combined seven benchmark sources including RareBench subsets, MyGene2, MIMIC-IV-Rare, DDD, and Phenopacket Store; many cases were public or publication-derived, and contamination was not excluded.
Zhao (DeepRare), 2026	GPT-4o standalone 7-benchmark	D5	Same seven public benchmarks as DeepRare; the standalone row removes case-bank augmentation but not plausible pre-training or public-benchmark exposure.
Chen X (RareAgents), 2024§	RareAgents	D2, D5, D7	GPT-4o judged whether outputs matched standard disease names without documented human validation; the memory module retrieved similar cases from RareBench itself; the fully equipped system was compared with less-equipped standalone baselines.
Chen X (RareAgents), 2024§	GPT-4o standalone RareBench	D2, D5, D7	Same unvalidated GPT-4o judge and RareBench public-benchmark exposure as the augmented RareAgents estimate; the standalone row lacks the memory-module pathway but shares the paper-level evaluation concerns.
Chen X (MAC), 2025	MAC multi-agent (GPT-4)	D6	Case selection after January 2022 reduced leakage risk, but the GPT-4 version was not pinned, limiting reproducibility and making an exact cutoff assessment impossible.
Rose (MEDDxAgent), 2025§	MEDDxAgent GPT-4o zero-shot	D5	Evaluation used a RareBench 100-case subset and a 102-disease candidate list; public-benchmark exposure and restricted-answer-set design may inflate absolute R@1 relative to open-ended generation.
Rose (MEDDxAgent), 2025§	MEDDxAgent GPT-4o few-shot (Dyn_BAIL)	D5	Same RareBench subset and 102-disease candidate list as the zero-shot row, with dynamic few-shot examples added.
Zheng (Deep-DxSearch), 2025§	Deep-DxSearch rare	D5	Retrieval database contained evaluation benchmark sources including RareBench, MME, LIRICAL, and MIMIC-R; no explicit query-case exclusion mechanism was documented.
Yang (RareSeek), 2025§	RareSeek-R1 EHR-Internal	D1, D3, D5, D6	Same-institution EHR training and test setting lacked documented temporal or institutional separation; reference labels were clinical rather than uniformly molecular; code, model, and data were not fully available.
Wang (MINT), 2026§	MINT GMDb	D5	GMDb split held out cases but not diseases; external validation showed substantially stronger performance on diseases present in training than on disjoint diseases.
Schumacher (RareScale), 2025§	RareScale augmented GPT-4o	D3, D5, D6	Reference standard was an expert-system differential rather than a confirmed clinical diagnosis; candidate generation and evaluation used the same simulated or expert-system disease source without disease-level holdout; GPT-4o version and data were not pinned.
Schumacher (RareScale), 2025§	RareScale baseline GPT-4o	D3, D5, D6	Same expert-system differential reference standard, simulated or expert-system disease source, and missing GPT-4o version and data details as the augmented RareScale estimate.
Chen X (RareBench), 2024	RareBench GPT-4	D5, D6	RareBench Task 4 combined public-benchmark-derived cases with PUMCH cases; 51% were public benchmark components predating the approximate cutoff of GPT-4 1106-preview; reproducibility was limited by partially restricted data and deprecated model and version details.
Aldin et al., 2026§	MIMIC-RD Llama-3.3	D3, D5	Reference diseases were mined from MIMIC-IV notes rather than molecularly confirmed; underlying MIMIC-IV notes were public or credentialed and dated 2008–2022, predating the Llama 3.3 cutoff.

Table S4 continued

Study	Selected meta-analysis entry	High-risk domains	Short rationale
Chimirri et al., 2025	GPT-4o English Phenopacket Store	D5, D6	Phenopacket Store cases were derived from hundreds of published case reports, and the source authors acknowledged possible model access to some source clinical data; GPT-4o was run at temperature 1 with single-run reporting.
Ilić et al., 2025	GPT-4-turbo skeletal dysplasia	D5, D6	The 45-patient cohort was novel and genetically confirmed, but skeletal-dysplasia phenotype–diagnosis associations are well represented in the literature; GPT-4-turbo version details were approximate.
Ao et al., 2025	Claude 3.5 Sonnet Chinese case repository	D5, D6	Chinese Medical Case Repository cases were public case reports likely predating the model cutoff; exact model versions, parameters, and run conditions were not fully reported.
Reese et al., 2026	o1-preview Phenopacket Store	D5	Phenopacket Store cases were derived from published case reports; benchmark-level data-leakage concern applies.
Olmo (DxGPT), 2024§	DxGPT Claude 3 Opus RAMEDIS	D2, D5, D6	GPT-4 0613 judged whether model outputs matched reference diagnoses without reported human validation; RAMEDIS is a long-standing public rare-disease benchmark; exact model versions and run settings were not fully reproducible.

§ = preprint (not peer-reviewed at time of search). Domains follow the modified QUADAS-3 instrument used in this review: D1 participant selection; D2 index test conduct; D3 target condition and reference standard; D4 analysis; D5 data leakage or contamination; D6 reproducibility; D7 evaluation fairness. Data leakage (D5) was the most common high-risk domain (18 of 19 entries), but high overall risk-of-bias was not solely attributable to leakage: several entries were rated high because output matching used unvalidated LLM judges, reference standards were simulated or weakly verified, model versions or run conditions were not reproducible, or comparators were not evaluated under symmetric conditions. Row-level data-leakage and model-cutoff evidence is provided in Supplementary Table S3.

Supplementary Table S5. Reviewer-requested sensitivity analyses for the pooled R@1 estimate, including alternative random-effects estimators, exclusion of small studies, exclusion of the Phenopacket Store benchmark, and restriction to peer-reviewed publications.

Analysis	k	N	Pooled R@1	95% CI	95% PI	I ²
Primary pooled estimate (DerSimonian–Laird random-effects)	19	39,529	43.2%	35.1%–51.6%	11.6%–78.3%	99.6%
Alternative random-effects estimator (REML)	19	39,529	43.2%	35.2%–51.5%	11.9%–78.0%	99.6%
Alternative random-effects estimator (REML + Hartung–Knapp)	19	39,529	43.2%	34.6%–52.1%	10.2%–80.2%	99.6%
Excluding small studies (N < 100)	18	39,484	42.7%	34.3%–51.2%	11.3%–77.9%	99.6%
Excluding small studies (N ≤ 100)	16	39,284	40.0%	31.5%–48.9%	9.7%–75.5%	99.7%
Excluding Phenopacket Store benchmark	17	29,399	46.0%	37.7%–54.5%	14.6%–79.4%	99.5%
Peer-reviewed publications only	12	30,672	40.8%	32.5%–49.4%	14.3%–70.6%	99.5%

k = number of system–dataset entries in the pool; N = total number of cases; CI = confidence interval; PI = prediction interval; DL = DerSimonian–Laird random-effects estimator; REML = restricted maximum likelihood estimator; Hartung–Knapp = Hartung–Knapp variance adjustment to the REML pooled estimate. The N < 100 analysis excludes one study (skeletal dysplasia, N=45). The N ≤ 100 analysis additionally excludes the two MEDDxAgent N=100 RareBench entries. For the peer-reviewed-only analysis, publication status was checked at extraction rather than inferred from identifier prefix alone. Across all sensitivity analyses, between-study heterogeneity remained extreme (I² > 99%) and the prediction interval remained wide, supporting the descriptive interpretation of the pooled estimate.

Supplementary Table S6. Diagnosis-output scoring approach and disease-label equivalence handling for the 15 included studies.

Study	Scoring approach	Synonym / ontology / equivalent-label handling
Zhao (DeepRare), 2026	LLM-as-judge (validated against specialists)	GPT-4o judged whether each top-5 predicted disease matched the gold-standard diagnosis (Prompt 2 in the source paper); reliability validated against 8 rare-disease specialists (Pearson $r=0.8689$, 88% concordance over 240 cases). Upstream, a BioLORD disease-name normalizer mapped predictions to Orphanet/OMIM identifiers, which was used to expand the gold-standard set with synonymous identifiers before the LLM judge was applied. The same judge framework was applied to the GPT-4o standalone comparator.
Chen X (RareAgents), 2024§	LLM-as-judge	GPT-4o judged whether model disease-name outputs matched the reference disease names; no human validation of the judge was reported. The same judge framework was applied to the GPT-4o standalone comparator.
Chen X (MAC), 2025	LLM-as-judge (validated against humans)	GPT-4o evaluator validated against a human panel; reported Fleiss' kappa 0.859–0.958.
Rose (MEDDxAgent), 2025§	Exact / constrained-label matching	Evaluation used a 102-disease candidate list derived from a RareBench subset; prompt required exact disease terminology. The same setup applied to both zero-shot and few-shot (Dyn_BAI) comparators.
Zheng (Deep-DxSearch), 2025§	Algorithmic; synonym detail insufficiently reported	Diagnostic-accuracy scoring used automated matching; a separate LLM judge was used for reasoning-quality evaluation and was reported as physician-validated; the synonym/ontology handling for diagnostic label matching was less explicitly reported.
Yang (RareSeek), 2025§	Algorithmic ontology mapping with hierarchical matching	Models generated up to 20 diagnoses normalized to Orphanet identifiers; both exact and hierarchical matching were reported.
Wang (MINT), 2026§	Exact / fuzzy benchmark-label matching	Automated matching against GMDb disease labels using string-similarity (SequenceMatcher threshold) comparison.
Schumacher (RareScale), 2025§	LLM-as-judge	GPT-4o binary judge for Top-1/Top-5 with a separate 5-level similarity judge classifying exact, relevant, somewhat relevant, and unrelated matches. The same judge framework was applied to both the augmented and baseline GPT-4o comparators.
Chen X (RareBench), 2024	Manual / benchmark-label verification	Ranked lists were manually verified against known diagnoses; source benchmark diagnoses include OMIM/Orphanet-coded public components and PUMCH clinical diagnoses.
Aldin et al., 2026§	Algorithmic ontology mapping	Reference diseases were mapped to Orphanet; Hit@k was evaluated by Orphanet code matching.
Chimirri et al., 2025	Algorithmic ontology mapping (subtype-aware)	pheval.llm parsed free-text diagnoses to Mondo identifiers; ontology-aware scoring counted disease subtype equivalence (e.g., related Loeys–Dietz syndrome subtypes).
Ilić et al., 2025	Manual / expert adjudication	A specialist in medical genetics adjudicated results; subtype discrepancies were accepted if the general diagnostic category was correct.
Ao et al., 2025	Insufficiently reported	Source publication does not provide enough detail to determine how synonyms or ontology normalization were handled.
Reese et al., 2026	Algorithmic ontology mapping (subtype-aware)	MALCO mapped LLM outputs to Mondo disease ontology terms via OAK/OntoGPT, then mapped Mondo to OMIM identifiers; disease grouping handled subtype variation.
Olmo (DxGPT), 2024§	LLM-as-judge	GPT-4 0613 judged whether each model's ranked diagnosis matched the reference label; no human validation of the judge was reported.

§ = preprint (not peer-reviewed at time of search). Where a study contributed both a baseline and an augmented entry to the meta-analysis, both arms used the same scoring procedure; this table reports the per-study procedure rather than repeating identical rows. Across the 15 included studies, source studies used heterogeneous scoring approaches: algorithmic non-judge matching (ontology or identifier mapping, exact or fuzzy benchmark-label matching) in 6 studies; LLM-as-judge evaluation in 5 studies; manual or expert adjudication in 2 studies; and insufficiently reported synonym or normalization details in 2 studies. Because case-level model outputs were generally unavailable, R@1 was extracted after applying each source study's own scoring procedure rather than re-scored under a common cross-study synonym dictionary. This evaluation-method heterogeneity is one contributor to between-study heterogeneity in pooled R@1.

PRISMA-DTA checklist

The PRISMA-DTA checklist below reports where each item is addressed in the manuscript.

Section/topic	#	PRISMA-DTA Checklist Item	Reported on page #
TITLE / ABSTRACT			
Title	1	Identify the report as a systematic review (+/- meta-analysis) of diagnostic test accuracy (DTA) studies.	Title page (p. 1)
Abstract	2	Abstract: See PRISMA-DTA for abstracts.	Abstract (p. 1)
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of what is already known.	Introduction (p. 2)
Clinical role of index test	D1	State the scientific and clinical background, including the intended use and clinical role of the index test, and if applicable, the rationale for minimally acceptable test accuracy (or minimum difference in accuracy for comparative design).	Introduction, paragraphs 2–3 (p. 2); Methods (p. 9)
Objectives	4	Provide an explicit statement of question(s) being addressed in terms of participants, index test(s), and target condition(s).	Introduction, paragraph 4 — three objectives (p. 2)
METHODS			
Protocol and registration	5	Indicate if a review protocol exists, if and where it can be accessed (e.g., Web address), and, if available, provide registration information including registration number.	Methods (p. 9); OSF osf.io/zebn5 ; DOI: 10.17605/OSF.IO/FMABN
Eligibility criteria	6	Specify study characteristics (participants, setting, index test(s), reference standard(s), target condition(s), and study design) and report characteristics (e.g., years considered, language, publication status) used as criteria for eligibility, giving rationale.	Methods (p. 9)
Information sources	7	Describe all information sources (e.g., databases with dates of coverage, contact with study authors to identify additional studies) in the search and date last searched.	Methods (p. 9); Supplementary Appendix
Search	8	Present full search strategies for all electronic databases and other sources searched, including any limits used, such that they could be repeated.	Supplementary Appendix (full search strategies)
Study selection	9	State the process for selecting studies (i.e., screening, eligibility, included in systematic review, and, if applicable, included in the meta-analysis).	Methods (p. 9); Figure 1 legend (p. 16)
Data collection process	10	Describe method of data extraction from reports (e.g., piloted forms, independently, in duplicate) and any processes for obtaining and confirming data from investigators.	Methods (p. 9)
Definitions for data extraction	11	Provide definitions used in data extraction and classifications of target condition(s), index test(s), reference standard(s) and other characteristics (e.g. study design, clinical setting).	Methods (p. 9)
Risk of bias and applicability	12	Describe methods used for assessing risk of bias in individual studies and concerns regarding the applicability to the review question.	Methods (p. 10); Supplementary Appendix (modified QUADAS-3 instrument)
Diagnostic accuracy measures	13	State the principal diagnostic accuracy measure(s) reported (e.g. sensitivity, specificity) and state the unit of assessment (e.g. per-patient, per-lesion).	Methods (p. 10) — R@1 as primary measure, per-case unit

Section/topic	#	PRISMA-DTA Checklist Item	Reported on page #
Synthesis of results	14	Describe methods of handling data, combining results of studies and describing variability between studies. This could include, but is not limited to: a) handling of multiple definitions of target condition; b) handling of multiple thresholds of test positivity; c) handling multiple index test readers; d) handling of indeterminate test results; e) grouping and comparing tests; f) handling of different reference standards.	Methods (p. 10)
Meta-analysis	D2	Report the statistical methods used for meta-analyses, if performed.	Methods (p. 10) — Freeman–Tukey double arcsine + DerSimonian–Laird random-effects
Additional analyses	16	Describe methods of additional analyses (e.g., sensitivity or subgroup analyses, meta-regression), if done, indicating which were pre-specified.	Methods (p. 10) — pre-specified subgroup analyses (augmentation, input modality); post-hoc exploratory benchmark analysis; sensitivity analyses
RESULTS			
Study selection	17	Provide numbers of studies screened, assessed for eligibility, included in the review (and included in meta-analysis, if applicable) with reasons for exclusions at each stage, ideally with a flow diagram.	Results: Study Selection and Characteristics (p. 3); Figure 1 legend (p. 16)
Study characteristics	18	For each included study provide citations and present key characteristics including: a) participant characteristics (presentation, prior testing); b) clinical setting; c) study design; d) target condition definition; e) index test; f) reference standard; g) sample size; h) funding sources.	Results (p. 3); Table 1 (p. 3)
Risk of bias and applicability	19	Present evaluation of risk of bias and concerns regarding applicability for each study.	Results: Risk of Bias (p. 5); Table 3 (p. 6)
Results of individual studies	20	For each analysis in each study (e.g. unique combination of index test, reference standard, and positivity threshold) report 2x2 data (TP, FP, FN, TN) with estimates of diagnostic accuracy and confidence intervals, ideally with a forest or receiver operator characteristic (ROC) plot.	Results: Overall Diagnostic Accuracy (p. 3); Figure 2 legend (p. 16)
Synthesis of results	21	Describe test accuracy, including variability; if meta-analysis was done, include results and confidence intervals.	Results: Overall Diagnostic Accuracy (p. 3); pooled R@1 43.3% (95% CI 35.1–51.6; $I^2=99.6\%$)
Additional analysis	23	Give results of additional analyses, if done (e.g., sensitivity or subgroup analyses, meta-regression; analysis of index test: failure rates, proportion of inconclusive results, adverse events).	Results: LLM Augmentation (p. 5); Benchmark Composition (p. 4); Input Modality (p. 5); Figure 3 legend (p. 16); Table 2 (p. 5); Figure S1 (supplementary)
DISCUSSION			
Summary of evidence	24	Summarize the main findings including the strength of evidence.	Discussion, paragraphs 1–2 (pp. 6–7)
Limitations	25	Discuss limitations from included studies (e.g. risk of bias and concerns regarding applicability) and from the review process (e.g. incomplete retrieval of identified research).	Discussion, limitations paragraph (p. 8)
Conclusions	26	Provide a general interpretation of the results in the context of other evidence. Discuss implications for future research and clinical practice (e.g. the intended use and clinical role of the index test).	Discussion, paragraphs 2–3 and closing (pp. 6–8) — implications for benchmarking, augmentation, and clinical validation
FUNDING			
Funding	27	For the systematic review, describe the sources of funding and other support and the role of the funders.	Acknowledgements (p. 10); Author Contributions (pp. 10–11)

Adapted from: McInnes MDF, Moher D, Thombs BD, McGrath TA, Bossuyt PM, The PRISMA-DTA Group (2018). Preferred Reporting Items for a Systematic Review and Meta-analysis of Diagnostic Test Accuracy Studies: The PRISMA-DTA Statement. *JAMA*. 2018 Jan 23;319(4):388–396. doi: 10.1001/jama.2017.19163.