arXiv:2607.24878v1 [q-bio.QM] 27 Jul 2026

GraphRareBench: An Auditable Graph-Evidence Benchmark for Phenotype-Driven Rare-Disease Diagnosis

Guiling Guo[1], **Jia Yang**[2], **Jiahao Xu**[3], **Shuyuan Zheng**[3], **Zhonghai Sun**[4], **Qiyuan Li**[3,1]

[1]National Institute for Data Science in Health and Medicine, Xiamen University
[2]Institute of Artificial Intelligence, Xiamen University
[3]School of Medicine, Xiamen University
[4]The First Affiliated Hospital of Xiamen University

Abstract

Phenotype-driven diagnostic benchmarks usually report the rank of the reference disease, but they rarely reveal which plausible alternatives are ranked above it or what evidence a tool-using model examines before making its decision. We introduce GraphRareBench, a provenance-preserving benchmark containing 2,365 ontology-derived cases and 18,093 target–confounder pairs. Each case includes a coarsened HPO query, a fixed candidate pool, graph-defined hard confounders, and source-linked evidence records. On the 237-case gene-component-disjoint test split, supervised rankers using a shared 21-feature interface achieved MRRs of 0.640–0.740 and case-averaged target-over-confounder accuracies of 0.898–0.916. Agents instantiated with Agents-A1 and DeepSeek-V4-Flash achieved MRRs of 0.746 and 0.718, respectively. Their paired MRR difference was not statistically significant, whereas their target-evidence coverage differed by 0.561. Together with the observation that 22.1%–43.7% of selected Hit@10 successes still ranked at least one graph-defined hard confounder above the target, these results indicate that full-pool retrieval, hard-confounder discrimination, and observable evidence access capture complementary aspects of model behavior. GraphRareBench therefore provides a foundation for more transparent and evidence-aware evaluation of phenotype-driven diagnostic systems. Code and dataset are avaliable at https://github.com/GUI0609/GraphRareBench.

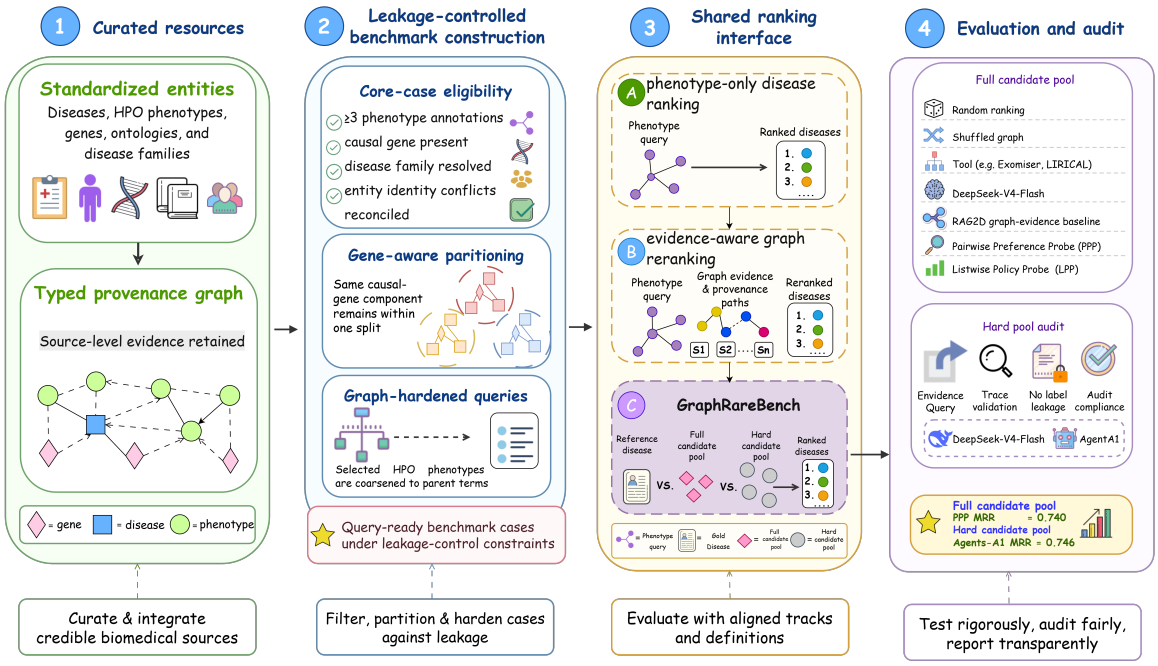

Figure 1: GraphRareBench converts curated rare-disease resources into ontology-derived ranking cases with fixed candidate pools, graph-defined hard alternatives, and source-linked evidence records. Evaluated systems receive the declared query, candidate pool, and evidence view; evaluator-only labels support target-over-confounder, mechanism-sliced, and tool-trace audits. Icons are from Icons8.com.

Introduction

Rare diseases are individually uncommon but collectively impose a substantial diagnostic burden. Many patients undergo prolonged diagnostic odysseys because early clinical presentations are heterogeneous, incomplete, and shared across multiple disorders. Although molecular testing has become central to confirming many genetic rare diseases, phenotype-driven reasoning remains essential for generating diagnostic hypotheses, selecting tests, prioritizing variants, guiding referrals, and constructing differential diagnoses (Manickam et al. 2021). Computational diagnostic systems must therefore address a fundamental clinical task: ranking candidate rare diseases from a compact and often incomplete phenotype description.

This task is inherently difficult because a small set of observed phenotypes may support several closely related diseases. Candidate diseases may share HPO terms, ontology ancestors, disease families, semantic neighborhoods, gene families, causal-gene components, or even the same causal gene. Standard top-k metrics capture whether the target disease appears near the top of the list, but not whether a clinically plausible alternative is ranked above it. A model may therefore achieve a favorable Hit@k score while still failing at the central differential-diagnosis problem. Meaningful evaluation must measure both target retrieval and the model's ability to distinguish the target from closely related alternatives.

This limitation becomes more consequential for large language model (LLM)-based diagnostic systems. LLMs can generate fluent and medically plausible rankings, yet the final output alone does not reveal whether the ranking is supported by case-specific evidence or by broad disease associations and prior knowledge. Moreover, final-ranking metrics alone may fail to reveal substantial differences in the evidence-seeking trajectories of tool-using systems. Evaluating these systems therefore requires more than a reference-disease rank: the candidate set must contain difficult alternatives, the evidence available for each candidate must remain traceable, and the model's evidence-access behavior must be auditable.

We introduce GraphRareBench, a provenance-preserving benchmark for phenotype-driven rare-disease ranking. Each case contains a coarsened HPO query, a fixed full candidate pool, a graph-defined subset of hard confounders, and evidence records linked to their original sources. Target and confounder annotations are retained by the evaluator, while evaluated methods receive only the declared query, candidate set, and permitted evidence interface. This design supports two complementary evaluations: full-pool ranking measures whether a method retrieves the target disease among a broad set of candidates, whereas a tool-mediated hard-pool audit examines how an agent distinguishes the target from its closest alternatives and which evidence it accesses before producing the final ranking.

Contributions. Our contributions are threefold: (i) we introduce GraphRareBench, comprising 2,365 ontology-derived cases and 18,093 target–confounder pairs, with coarsened HPO queries, fixed candidate pools, gene-component-aware data partitions, seven graph-defined confounder mechanisms, and source-linked evidence records; (ii) we provide target-over-confounder metrics and mechanism-specific analyses that directly evaluate whether a model can distinguish the target disease from closely related alternatives; and (iii) we establish a unified evaluation framework for both final ranking performance and observable evidence-access behavior, covering phenotype-driven tools, prompted LLMs, supervised graph-evidence rankers, and tool-using agents.

Related Work

Rare-disease resources and phenotype-driven tools. Phenotype-driven rare-disease diagnosis relies on standardized vocabularies, disease ontologies, molecular databases, and knowledge graphs that connect clinical phenotypes with diseases, genes, variants, and diagnostic panels. HPO, Orphanet, OMIM, and Mondo provide complementary representations of phenotypes and rare diseases, while GenCC, ClinGen, ClinVar, PanelApp, and KGRD contribute gene–disease validity, variant interpretation, diagnostic-panel information, and graph-structured biomedical evidence (Ma, NM et al. 2024; Rath et al. 2012; Amberger et al. 2015; Vasilevsky et al. 2026; DiStefano et al. 2022; Andersen et al. 2025; Landrum et al. 2025; Martin et al. 2019; Guo et al. 2026). Phenotype-driven systems such as Phenomizer, Exomiser, LIRICAL, Phen2Gene, and PhenoBrain use these resources to prioritize candidate diseases, genes, or variants from patient phenotypes (Köhler et al. 2009; Smedley et al. 2015; Robinson et al. 2020; Zhao et al. 2020; Mao et al. 2025).

Tool-using medical agents and trace evaluation. Tool-using medical agents turn rare-disease diagnosis into a sequence of information-processing and evidence-integration steps rather than a single free-form prediction. RareAgents combines multidisciplinary-team coordination, long-term memory, and specialized phenotype and treatment-oriented tools, allowing different physician agents to retrieve similar cases, query diagnostic systems, exchange findings, and synthesize a final report (Chen et al. 2026b). DeepRare extends this pattern with a central coordinating agent and more than 40 specialized tools and knowledge sources for phenotype normalization, disease normalization, case retrieval, phenotype and genotype analysis, literature search, and evidence-linked differential diagnosis (Zhao et al. 2026). KGRD adds knowledge-graph-guided gene and phenotype inference, patient-level case retrieval, multidisciplinary specialist deliberation, and a verifier that checks candidate support across multiple evidence channels (Guo et al. 2026). These systems therefore expose richer intermediate artifacts than one-shot LLMs, including retrieved cases, tool outputs, agent discussions, graph paths, and source-linked evidence. Such artifacts make it possible to ask not only whether the final diagnosis is correct, but also which evidence channels were selected, whether alternatives were examined, and how the final ranking was assembled.

Rare-disease evaluation benchmarks. Phenopacket Store and PhEval provide standardized case representations and reproducible evaluation protocols, while RareBench, RareArena, MIMIC-RD, UDN-derived studies, RareSyn, and SHEPHERD

extend benchmarking across clinical cases, language-model diagnosis, and graph-based systems (Danis et al. 2025; Bridges et al. 2025; Chen et al. 2024, 2026a; AlDin et al. 2025; Shyr et al. 2025; Wang et al. 2025; Alsentzer et al. 2025). Collectively, these resources have strengthened the evaluation of target-disease retrieval, but their primary outcomes remain rank-based measures such as reciprocal rank, top-k retrieval, or candidate-ranking accuracy. Large-scale comparisons further suggest that, for phenotype-only diagnosis, LLMs remain less reliable than established rare-disease decision-support tools (Reese et al. 2026). In parallel, general agent benchmarks such as AgentBench, ToolSandbox, and τ-bench emphasize that tool-use validity, intermediate actions, and task completion should be evaluated in addition to final answers (Liu et al. 2024; Lu et al. 2025; Yao et al. 2024). However, existing medical and agent benchmarks rarely connect observable action traces to a fixed differential-diagnosis candidate set containing explicitly labeled hard confounders and source-traceable evidence. GraphRareBench addresses this gap by fixing the candidate pool, identifying hard alternatives through graph relations, and jointly auditing the final ranking and the evidence requested during tool use.

Task Formulation

GraphRareBench formulates phenotype-driven rare-disease ranking as a closed candidate-set task. Each case i is derived from an ontology-linked disease profile and contains one target disease together with a set of candidate alternatives. The input visible to an evaluated method is

$$X_i = (Q_i, C_i, E_i),$$

where Q_i is the coarsened phenotype query, C_i is the candidate set used in the current evaluation, and E_i is the evidence view permitted by the corresponding interface. The target disease $g_i \in C_i$ is hidden from the method and used only for evaluator-side scoring.

GraphRareBench provides two candidate-set configurations. The full candidate pool is

$$C_i^{\text{full}} = \{g_i\} \cup H_i \cup A_i,$$

where H_i contains graph-defined hard confounders and A_i contains additional candidate diseases. This pool evaluates whether a method can retrieve the target from a broader set of alternatives. The separate tool-mediated audit uses the hard candidate pool

$$C_i^{\text{hard}} = \{g_i\} \cup H_i,$$

which focuses on distinguishing the target from its closest graph-defined alternatives.

Each evaluated method must return a complete ranking of all diseases in the active candidate set C_i. Target identities, hard-confounder annotations, and confounder-reason labels are never exposed through the evaluation interface. Let $\mathrm{rank}_i(c)$ denote the one-indexed position of candidate c in the returned ranking.

Metrics

We report standard MRR and Hit@k for target retrieval. Let $\mathrm{rank}_i(c)$ denote the one-indexed rank of candidate c. For cases containing hard confounders, target-over-confounder accuracy is defined as

$$\mathrm{ToC}_i = \frac{1}{|H_i|} \sum_{h \in H_i} \mathbb{I}[\mathrm{rank}_i(g_i) < \mathrm{rank}_i(h)],$$

$$\mathrm{ToC}_{\text{case}} = \frac{1}{N_C} \sum_{i:\,|H_i|>0} \mathrm{ToC}_i,$$

$$\mathrm{ToC}_{\text{pair}} = \frac{\sum_{i:\,|H_i|>0} |H_i| \mathrm{ToC}_i}{\sum_{i:\,|H_i|>0} |H_i|},$$

where N_C is the number of cases with at least one hard confounder. Thus, $\mathrm{ToC}_{\text{case}}$ weights cases equally, whereas $\mathrm{ToC}_{\text{pair}}$ weights target–confounder pairs equally. We additionally report $\mathrm{AnyConfAbove}$, the proportion of cases with $\mathrm{ToC}_i < 1$. Mechanism-specific analyses compute $\mathrm{ToC}_{\text{pair}}$ after restricting comparisons to pairs carrying the corresponding confounder label.

GraphRareBench Construction

Resource harmonization. GraphRareBench is built from normalized rare-disease records that link disease identities to HPO phenotype annotations, causal genes, ontology relations, disease-family metadata, and source-level evidence. Starting from 7,441 normalized disease records, we retained 2,371 diseases that satisfied the predefined requirements for phenotype coverage, gene annotation, family metadata, and identity resolution. A final label-leakage screen removed six additional diseases, resulting in 2,365 released cases. Each retained disease was required to have at least three disease-level HPO annotations. The resulting graph contains 10,754 phenotype nodes, 4,262 gene nodes, 212,901 typed modeling edges, and 515,774 source assertions. We additionally derive a smaller model-visible evidence layer that supports controlled evidence access and provenance auditing.

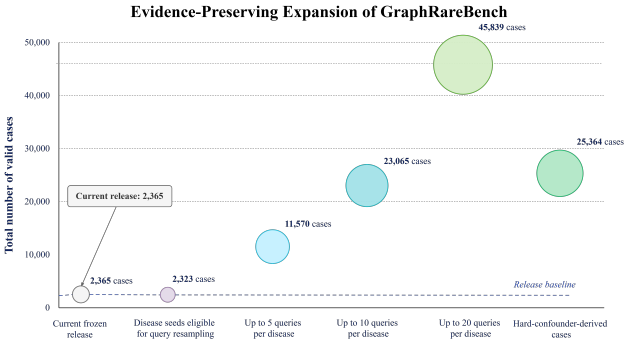

Figure 2: Frozen-release scale and controlled expansion capacity. The current benchmark contains 2,365 fixed cases, of which 2,323 diseases are eligible for additional query construction. At query caps of 5, 10, and 20 per eligible disease, deterministic validity and deduplication filters yield 11,570, 23,065 and 45,839 disease-query cases, respectively. The hard-confounder-derived route yields 25,364 distinct disease-query cases.

Gene-aware partitioning. To reduce information leakage across genetically related diseases, GraphRareBench uses causal-gene annotations when constructing the train, validation, and test partitions. Diseases connected through the same causal-gene component are assigned to the same partition whenever gene annotations are available. This procedure produces 1,892 training cases, 236 validation cases, and 237 test cases. Candidate mining is performed only among diseases assigned to the same partition as the target. Evidence records associated with test diseases remain available through the declared benchmark interfaces. The resulting setting therefore evaluates whether supervised rankers generalize to target diseases from causal-gene components not observed during training, while preserving a shared evidence-access environment at test time.

Query abstraction and evidence preservation. For each eligible target disease g_i, GraphRareBench constructs a model-visible phenotype query Q_i together with a source-traceable evidence record. Across the 2,365 released cases, the visible queries contain 9,423 HPO terms, with a median of four terms per case and a range of two to four. Of these terms, 9,208 (97.7%) are coarsened ontology parents, whereas 215 (2.3%) are retained directly. All mappings were validated against the registered HPO OBO release dated 2026-02-16. Each coarsened term is exactly one upward ontology step from its source phenotype, while each direct term has distance zero. No case contains duplicate visible HPO identifiers.

The coarsening procedure modifies only the phenotype query presented to the model. The original source phenotypes are preserved in the provenance records and may be accessed only through the evidence interfaces permitted in the corresponding evaluation setting. This design reduces direct phenotype matching while retaining the evidence required for controlled retrieval and audit.

Candidate-pool construction and annotations. Each case uses the two fixed candidate views defined above. Hard confounders are mined deterministically within each partition using high HPO overlap, SapBERT semantic neighbors (Liu et al. 2021), shared direct MONDO parents, disease-family membership, shared causal-gene families, shared causal-gene components, and exact causal-gene sharing. Because these mechanisms overlap, each target–confounder pair may receive multiple labels. Additional non-hard candidates are drawn from semantic, phenotype, graph-neighbor, and deterministic random sources after labeled hard confounders are excluded. Candidate order is independently permuted, and all evaluator-side labels remain hidden at test time. The frozen release contains 18,093 labeled target–confounder pairs; detailed frequencies, overlaps, thresholds, and candidate-cap analyses are reported in the supplementary material.

Experiments

We organize the experiments around four questions:
Q1: Metric complementarity. Do full-pool retrieval metrics and target-over-confounder metrics reveal different ranking failures?
Q2: Interface capacity. How much ranking performance is achieved by methods with no GraphRareBench supervision, prompted LLMs, and supervised graph-evidence rankers?
Q3: Confounder mechanisms. Which types of graph-defined hard confounders remain difficult to distinguish from the target?
Q4: Evidence and trace audit. What can fixed-feature interventions and tool-use traces reveal about the evidence used to produce a ranking?

Experimental Protocol

All full-pool methods rank the same fixed candidate set for each case. At test time, the target identity, hard-confounder labels, candidate-construction roles, graph-derived ranks and scores, and other evaluator-only fields are hidden unless they are explicitly

included in the evaluated interface.

The gene-component-aware split contains 1,892 training cases, 236 validation cases, and 237 held-out test cases. On the test partition, each case contains a median of 69 full-pool candidates, including a median of five hard confounders. We report full-pool MRR, Hit@1, Hit@5, Hit@10, and case-averaged target-over-confounder accuracy. Analyses by confounder mechanism use pair-averaged target-over-confounder accuracy. Confidence intervals are estimated through case-level bootstrap resampling. Paired bootstrap resampling is used for method contrasts evaluated on the same cases.

LLM and agent results are obtained from one fixed protocol run. Accordingly, these results characterize behavior under the reported prompts, model versions, decoding settings, and candidate orders rather than variation across repeated model runs. Display-order analyses are performed on the outputs of these fixed runs.

Compared methods. Random ranking and shuffled-graph ranking serve as negative controls. HPO Similarity, Exomiser, and LIRICAL are phenotype-driven methods that use no GraphRareBench training labels. DeepSeek-V4-Flash (DeepSeek-AI 2026) is evaluated under three input conditions: coarsened HPO terms, source-support HPO terms, and compact candidate evidence cards. RAG2D, PPP, and LPP are trained on the GraphRareBench training and validation partitions using the same 21-feature interface. RAG2D provides a linear reference, whereas PPP and LPP provide nonlinear pairwise and listwise ranking models over the same 21-feature interface.

Q1–Q2: Full Candidate-Pool Ranking

Full-pool retrieval and hard-confounder ordering capture different failures. Table 1 reports results on the held-out test partition. Random ranking yields an MRR of 0.070, while the shuffled-graph control yields 0.077. HPO Similarity achieves a relatively low full-pool MRR of 0.170 but $\mathrm{ToC}_{\mathrm{case}}$ of 0.887. It therefore often ranks the target above the labeled hard confounders even when many additional candidates remain above the target in the full pool. This contrast shows that target retrieval and discrimination against the hard-confounder subset measure different aspects of ranking performance.

To examine failures that remain hidden by Hit@10, we compute the proportion of Hit@10 cases in which at least one hard confounder is still ranked above the target. Even among cases counted as Hit@10 successes, at least one graph-defined hard confounder remained above the target in 43.7% of DeepSeek cases, 29.3% of RAG2D cases, and 22.1% of PPP cases. Thus, a target can satisfy the conventional Hit@10 criterion while remaining below a clinically plausible alternative.

Phenotype-driven tools. Among methods that use no GraphRareBench supervision, Exomiser and LIRICAL achieve MRRs of 0.322 and 0.502, respectively. Their native scores do not cover every candidate: Exomiser scores 74.6% of candidates and 97.9% of targets, while LIRICAL scores 80.2% of candidates and 94.9% of targets. Replacing the primary unscored-tail convention with tied average ranks changed MRR by less than 10^{-4}, indicating that the reported MRRs are insensitive to this specific tail-ranking convention.

LLM input conditions. DeepSeek-V4-Flash achieves an MRR of 0.247 when given the coarsened HPO query. Replacing the coarsened terms with the curator-derived source-support terms increases MRR to 0.313, corresponding to a paired improvement of 0.066 (95% CI: 0.032–0.102). Because these source terms provide the original disease-level phenotype support from which the visible query was constructed, this condition represents a more informative upper-bound input rather than the standard benchmark setting.

Providing compact evidence cards changes DeepSeek's MRR from 0.247 to 0.219, while increasing its case-averaged target-over-confounder accuracy from 0.686 to 0.772. The paired changes are $\Delta\mathrm{MRR} = -0.028$ (95% CI: -0.069–0.012) and $\Delta\mathrm{ToC}_{\mathrm{case}} = 0.086$ (95% CI: 0.039–0.133). In this fixed run, the evidence cards therefore improve ordering against hard confounders without improving the target's position in the full candidate pool. This result is specific to the reported evidence format and protocol; it does not establish a general effect of evidence augmentation or evaluate probability calibration.

Supervised graph-evidence rankers. Using GraphRareBench training and validation supervision, the linear RAG2D reranker achieves an MRR of 0.640, Hit@1 of 0.464, Hit@10 of 0.966, and $\mathrm{ToC}_{\mathrm{case}}$ of 0.898. Its paired MRR difference is 0.138 relative to LIRICAL (95% CI: 0.073–0.202), 0.393 relative to DeepSeek with coarsened HPO input (95% CI: 0.337–0.448), and 0.421 relative to the DeepSeek evidence-card condition (95% CI: 0.372–0.470). These comparisons show the level of performance attainable when the shared graph-evidence features are fitted using in-benchmark supervision. Because the compared methods differ in both supervision and information interface, the contrasts should be interpreted as interface-level performance comparisons rather than isolated estimates of model-architecture effects.

On the same 21-feature interface, the nonlinear PPP and LPP rankers achieve MRRs of 0.740 and 0.730, respectively. PPP and LPP achieve MRR point estimates that are 0.100 and 0.090 higher than RAG2D, respectively. Because the models differ in both functional form and ranking objective, these results suggest that higher-capacity ranking models can extract additional predictive signal from the shared feature interface.

Q3: Reason-Sliced Confounder Analysis

Figure 3 reports mechanism-specific $\mathrm{ToC}_{\mathrm{pair}}$; multi-label pairs contribute to every applicable, and therefore overlapping, mechanism slice. The supervised rankers perform particularly well on semantic-neighbor confounders: RAG2D achieves a

Method	Input interface	Supervision	MRR	Hit@1	Hit@5	Hit@10	$\mathrm{ToC}_{\mathrm{case}}$	MRR 95% CI
Negative controls								
Random Ranking	Candidates	None	0.070	0.013	0.059	0.173	0.532	[0.056, 0.088]
Shuffled Graph	Permuted disease–HPO profiles	None	0.077	0.021	0.093	0.139	0.488	[0.058, 0.098]
Zero-label phenotype tools								
HPO Similarity	Coarsened HPO	None	0.170	0.072	0.219	0.384	0.887	[0.139, 0.204]
Exomiser	Coarsened HPO	External tool	0.322	0.160	0.489	0.684	0.748	[0.282, 0.365]
LIRICAL	Coarsened HPO	External tool	0.502	0.359	0.692	0.793	0.774	[0.451, 0.552]
Prompt-only LLM								
DeepSeek-V4-Flash	Coarsened HPO + candidates	Prompt	0.247	0.110	0.376	0.570	0.686	[0.211, 0.286]
DeepSeek-V4-Flash	Source-support HPO + candidates	Prompt	0.313	0.160	0.477	0.654	0.711	[0.272, 0.357]
DeepSeek-V4-Flash	Coarsened HPO + evidence cards	Prompt	0.219	0.072	0.346	0.549	0.772	[0.187, 0.252]
Supervised train/dev graph-evidence rankers								
RAG2D linear reranker	21 graph-evidence features	Train/dev	0.640	0.464	0.899	0.966	0.898	[0.595, 0.686]
PPP nonlinear pairwise	21 graph-evidence features	Train/dev	**0.740**	**0.595**	**0.928**	**0.975**	**0.916**	**[0.697, 0.782]**
LPP nonlinear listwise	21 graph-evidence features	Train/dev	0.730	0.586	0.903	0.970	0.903	[0.687, 0.773]

Table 1: Full candidate-pool ranking results on the held-out test partition. All methods rank the same released C_i^{full}. MRR and Hit@k score target recovery, whereas $\mathrm{ToC}_{\mathrm{case}}$ scores target-over-hard-confounder ordering. RAG2D, PPP, and LPP share the same 21-feature interface and use GraphRareBench train/dev supervision; cross-block comparisons quantify performance across different information regimes.

PPP score-time variant	Feat.	MRR	Hit@1	Hit@5	Hit@10	$\mathrm{ToC}_{\mathrm{case}}$
Full PPP checkpoint	21	0.740	0.595	0.928	0.975	0.916
Case-specific evidence only	8	0.510	0.316	0.776	0.873	0.851
Disease-profile/context only	11	0.547	0.380	0.793	0.907	0.846
Aggregate graph scores only	2	0.124	0.008	0.169	0.371	0.719
Strict non-query profile/centrality only	6	0.024	0.000	0.000	0.000	0.565

Table 2: Score-time feature-channel isolation for the frozen PPP checkpoint. Each row retains only the indicated feature group while keeping training and model selection fixed. Results reflect checkpoint sensitivity rather than independent feature contributions.

$\mathrm{ToC}_{\mathrm{pair}}$ of 0.980, while PPP and LPP both reach 0.990. Performance is lower for disease-family confounders, with scores of 0.850 for RAG2D, 0.871 for PPP, and 0.843 for LPP. Gene-context confounders also reveal substantial differences between methods: Exomiser achieves 0.659, compared with 0.936 and 0.939 for PPP and LPP, respectively. These results show that strong aggregate ranking performance does not imply uniform discrimination across confounder types. Across the supervised rankers, error rates were consistently higher for disease-family and gene-context confounders than for semantic-neighbor confounders. 2,530 of 18,093 pairs (14.0%) carried more than one mechanism label. Confidence intervals for the reason-sliced results use case-clustered resampling so that multiple confounders and overlapping labels from the same case are not treated as independent observations.

Q4: Evidence-Channel and Agent-Trace Audit

GraphRareBench audits evidence through two complementary views. First, we apply score-time feature isolation to the frozen PPP checkpoint while keeping the evaluation cases, training, and model selection fixed. Recomputed scores reproduce the stored rankings, with a maximum difference of 1.53×10^{-5} and no rank mismatches. Because PPP is nonlinear, Table 2 measures retained-set sufficiency rather than independent feature contribution or necessity. With all 21 features, MRR is 0.740; the case-specific and disease-profile/context sets retain MRRs of 0.510 and 0.547, whereas the aggregate graph-score and strict non-query sets yield 0.124 and 0.024. Aggregate graph scores retain some target-over-confounder separation, while strict non-query features preserve little ranking signal. Second, the tool-mediated audit records the disease profiles and query-conditioned evidence requested before ranking.

We next evaluate tool-using agents on the hard-confounder pool. For each case, the agent receives a coarsened phenotype query and a shuffled list of candidate disease names and IDs. It may make four tool calls before returning a complete ranking of all candidates. Two tools are available: a disease-profile tool that retrieves general information about selected diseases and

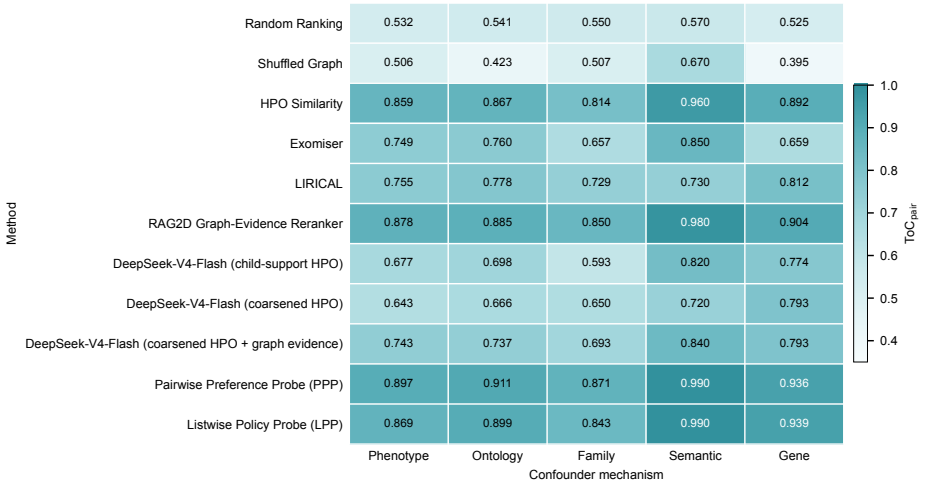

Figure 3: Reason-sliced target-over-confounder heatmap on the test partition. Entries are $\mathrm{ToC}_{\mathrm{pair}}$, the fraction of target–confounder pairs where the target disease is ranked above the hard confounder. Pairs with multiple reason labels contribute to each matching column. Slice case/pair counts are phenotype 237/474, ontology 186/338, family 98/140, semantic 94/100, and gene 148/314.

Metric	DeepSeek-V4-Flash	Agents-A1
MRR	0.718	0.746
Hit@1	0.540	0.582
$\mathrm{ToC}_{\mathrm{case}}$	0.817	0.833
Profile calls/case	1.89	3.51
Evidence calls/case	2.11	0.49
Candidates observed/call	2.58	2.34
Unique candidates/case	5.79	5.54
Tool switches/case	1.65	0.58
Target evidence coverage	0.937	0.376
Hard-confounder evidence coverage	0.863	0.254

Table 3: Tool-mediated hard-pool audit. Call and switch counts are averaged per case; candidate observations are averaged per tool call. Evidence coverage measures the proportion of targets or hard confounders for which query-conditioned evidence was retrieved.

a query-conditioned evidence tool that retrieves evidence linking selected candidates to the patient phenotype query. Each call can include up to three candidate diseases.

We evaluate agents instantiated with DeepSeek-V4-Flash and Agents-A1 (Bai et al. 2026). Deterministic checks verify that each run respects the permitted information boundary, makes only valid tool calls, and returns a complete candidate ranking. An independent LLM auditor additionally reviews compliance with the tool-use protocol. Ranking metrics are computed directly from the final candidate order.

As shown in Table 3, the Agents-A1-based agent achieves an MRR of 0.746, compared with 0.718 for the DeepSeek-based agent. The paired MRR difference is 0.029, with a 95% confidence interval of -0.010 to 0.068 and $p = 0.155$. Because the confidence interval includes zero, the observed difference in final ranking performance is not statistically significant.

The two agents nevertheless exhibit large descriptive differences in tool-use patterns in this fixed run. The DeepSeek-based agent makes an average of 2.11 query-conditioned evidence calls per case, whereas the Agents-A1-based agent makes 0.49. It also switches more frequently between the disease-profile and query-conditioned evidence tools, averaging 1.65 switches per case compared with 0.58 for the Agents-A1-based agent. Correspondingly, the DeepSeek-based agent retrieves evidence for a larger proportion of both target diseases and hard confounders. Its target-evidence coverage is higher by 0.561 (95% CI: 0.494–0.624), and its hard-confounder evidence coverage is higher by 0.609 (95% CI: 0.570–0.648). In this fixed protocol run, an MRR difference of 0.029 co-occurred with substantially different evidence-seeking patterns, including a target-evidence coverage difference of 0.561 and a hard-confounder evidence coverage difference of 0.609.

Discussion

GraphRareBench provides a controlled basis for evaluating modular diagnostic agents. A planning component can select candidates and evidence sources for further investigation, while specialized executors perform phenotype matching, graph traversal, gene-context retrieval, or source verification. Because candidate-level actions are aligned with hidden confounder annotations and source-linked evidence, the benchmark can separate gains arising from search policies, evidence executors, and allocation of a limited tool budget.

The fixed candidate pools also support controlled perturbations of phenotype granularity, candidate order, evidence availability, graph relations, source diversity, and tool-call budgets. Such interventions can distinguish failures caused by unavailable evidence from failures of evidence retrieval or integration. An open-world extension could first retrieve diseases from a larger knowledge base and then evaluate whether the target is correctly ordered against the hard alternatives in the retrieved set. Longitudinal phenotypes, uncertain or negated findings, laboratory measurements, and variant-level evidence provide further extensions.

Finally, the frozen candidate pools, evidence views, confounder mechanisms, and trace checks make GraphRareBench suitable as a regression suite. Updates to models, retrievers, or knowledge bases can be evaluated for aggregate gains and for localized regressions across confounder mechanisms, phenotype abstraction, query-independent shortcuts, and evidence-acquisition behavior.

Limitations

GraphRareBench evaluates ontology-derived cases within fixed candidate pools and therefore does not capture the full complexity of open-world clinical diagnosis, including noisy patient narratives, longitudinal phenotype evolution, uncertain or negated findings, and variant-level interpretation. Its hard confounders are defined through reproducible graph and semantic relations; although these relations provide controlled and challenging alternatives, they may not cover every distinction encountered in clinical practice.

Conclusion

We introduced GraphRareBench, a provenance-preserving benchmark for phenotype-driven rare-disease ranking. GraphRareBench combines coarsened phenotype queries, fixed candidate pools, graph-defined hard confounders, source-linked evidence, and hidden evaluator annotations within a unified evaluation framework. Its full-pool and target-over-confounder metrics reveal complementary ranking failures that conventional top-k evaluation can obscure. Feature-channel interventions and tool-mediated traces further distinguish final ranking performance from the evidence accessed to produce it. By jointly evaluating target retrieval, discrimination against plausible alternatives, and observable evidence-seeking behavior, GraphRareBench provides a reproducible test bed for developing and auditing evidence-aware rare-disease diagnostic systems.

References

AlDin, Z. E.; Wu, J.; Fung, J. P.; King, J.; Watts, M.; ONeill, L.; Cross, A. R.; and Sun, J. 2025. MIMIC-RD: Can LLMs differentially diagnose rare diseases in real-world clinical settings? *arXiv preprint arXiv:2601.11559*.

Alsentzer, E.; Li, M. M.; Kobren, S. N.; Noori, A.; Network, U. D.; Kohane, I. S.; and Zitnik, M. 2025. Few shot learning for phenotype-driven diagnosis of patients with rare genetic diseases. *NPJ digital medicine*, 8(1): 380.

Amberger, J. S.; Bocchini, C. A.; Schiettecatte, F.; Scott, A. F.; and Hamosh, A. 2015. OMIM. org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic acids research*, 43(D1): D789–D798.

Andersen, E. F.; Azzariti, D. R.; Babb, L.; Berg, J. S.; Biesecker, L. G.; Bly, Z.; Buchanan, A. H.; DiStefano, M. T.; Gong, L.; Harrison, S. M.; et al. 2025. The Clinical Genome Resource (ClinGen): advancing genomic knowledge through global curation. *Genetics in Medicine*, 27(1): 101228.

Bai, L.; Cao, Z.; Chen, Y.; Cui, Z.; Du, S.; Fan, Y.; Feng, S.; Guo, Z.; He, H.; He, L.; et al. 2026. Scaling the Horizon, Not the Parameters: Reaching Trillion-Parameter Performance with a 35B Agent. *arXiv preprint arXiv:2606.30616*.

Bridges, Y.; Souza, V. d.; Cortes, K. G.; Haendel, M.; Harris, N. L.; Korn, D. R.; Marinakis, N. M.; Matentzoglu, N.; McLaughlin, J. A.; Mungall, C. J.; et al. 2025. Towards a standard benchmark for phenotype-driven variant and gene prioritisation algorithms: PhEval-Phenotypic inference Evaluation framework. *BMC bioinformatics*, 26(1): 87.

Chen, H.; Zhao, Z.; Zhou, S.; Hu, S.; Wang, J.; Jin, Y.; Jin, X.; Tham, Y. C.; Wang, X.; Ma, W.; et al. 2026a. RareArena: a comprehensive benchmark dataset unveiling the potential of large language models in rare disease diagnosis. *The Lancet Digital Health*, 8(2).

Chen, X.; Jin, Y.; Mao, X.; Wang, L.; Zhang, S.; and Chen, T. 2026b. Rareagents: Autonomous multi-disciplinary team for rare disease diagnosis and treatment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, 101–109.

Chen, X.; Mao, X.; Guo, Q.; Wang, L.; Zhang, S.; and Chen, T. 2024. RareBench: can LLMs serve as rare diseases specialists? In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*, 4850–4861.

Danis, D.; Bamshad, M. J.; Bridges, Y.; Caballero-Oteyza, A.; Cacheiro, P.; Carmody, L. C.; Chimirri, L.; Chong, J. X.; Coleman, B.; Dalgleish, R.; et al. 2025. A corpus of GA4GH Phenopackets: case-level phenotyping for genomic diagnostics and discovery. *Human Genetics and Genomics Advances*, 6(1).

DeepSeek-AI. 2026. DeepSeek V4 Technical Documentation. Model card. Published April 27, 2026; accessed July 11, 2026.

DiStefano, M. T.; Goehringer, S.; Babb, L.; Alkuraya, F. S.; Amberger, J.; Amin, M.; Austin-Tse, C.; Balzotti, M.; Berg, J. S.; Birney, E.; et al. 2022. The Gene Curation Coalition: A global effort to harmonize gene–disease evidence resources. *Genetics in Medicine*, 24(8): 1732–1742.

Guo, G.; Shao, Z.; Luo, H.; Fu, Z.; Xiong, H.; Li, J.; Qin, Q.; Yang, X.; Hu, S.; Wu, J.; et al. 2026. KGRD: a knowledge-graph-augmented automated reasoning framework for diagnosis and counselling of paediatric rare genetic disorders. *npj Digital Medicine*.

Köhler, S.; Schulz, M. H.; Krawitz, P.; Bauer, S.; Dölken, S.; Ott, C. E.; Mundlos, C.; Horn, D.; Mundlos, S.; and Robinson, P. N. 2009. Clinical diagnostics in human genetics with semantic similarity searches in ontologies. *The American Journal of Human Genetics*, 85(4): 457–464.

Landrum, M. J.; Chitipiralla, S.; Kaur, K.; Brown, G.; Chen, C.; Hart, J.; Hoffman, D.; Jang, W.; Liu, C.; Maddipatla, Z.; et al. 2025. ClinVar: updates to support classifications of both germline and somatic variants. *Nucleic Acids Research*, 53(D1): D1313–D1321.

Liu, F.; Shareghi, E.; Meng, Z.; Basaldella, M.; and Collier, N. 2021. Self-alignment pretraining for biomedical entity representations. In *Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: human language technologies*, 4228–4238.

Liu, X.; Yu, H.; Zhang, H.; Xu, Y.; Lei, X.; Lai, H.; Gu, Y.; Ding, H.; Men, K.; Yang, K.; et al. 2024. Agentbench: Evaluating llms as agents. In *International Conference on Learning Representations*, volume 2024, 52989–53046.

Lu, J.; Holleis, T.; Zhang, Y.; Aumayer, B.; Nan, F.; Bai, H.; Ma, S.; Ma, S.; Li, M.; Yin, G.; et al. 2025. Toolsandbox: A stateful, conversational, interactive evaluation benchmark for llm tool use capabilities. In *Findings of the Association for Computational Linguistics: NAACL 2025*, 1160–1183.

Ma, G.; NM, B.; et al. 2024. The Human Phenotype Ontology in 2024: phenotypes around the world. *Nucleic Acids Res [Internet]*, 52: D1.

Manickam, K.; McClain, M. R.; Demmer, L. A.; Biswas, S.; Kearney, H. M.; Malinowski, J.; Massingham, L. J.; Miller, D.; Yu, T. W.; Hisama, F. M.; et al. 2021. Exome and genome sequencing for pediatric patients with congenital anomalies or intellectual disability: an evidence-based clinical guideline of the American College of Medical Genetics and Genomics (ACMG). *Genetics in Medicine*, 23(11): 2029–2037.

Mao, X.; Huang, Y.; Jin, Y.; Wang, L.; Chen, X.; Liu, H.; Yang, X.; Xu, H.; Luan, X.; Xiao, Y.; et al. 2025. A phenotype-based AI pipeline outperforms human experts in differentially diagnosing rare diseases using EHRs. *NPJ Digital Medicine*, 8(1): 68.

Martin, A. R.; Williams, E.; Foulger, R. E.; Leigh, S.; Daugherty, L. C.; Niblock, O.; Leong, I. U.; Smith, K. R.; Gerasimenko, O.; Haraldsdottir, E.; et al. 2019. PanelApp crowdsources expert knowledge to establish consensus diagnostic gene panels. *Nature genetics*, 51(11): 1560–1565.

Rath, A.; Olry, A.; Dhombres, F.; Brandt, M. M.; Urbero, B.; and Ayme, S. 2012. Representation of rare diseases in health information systems: the Orphanet approach to serve a wide range of end users. *Human mutation*, 33(5): 803–808.

Reese, J. T.; Chimirri, L.; Bridges, Y.; Danis, D.; Caufield, J. H.; Gargano, M. A.; Kroll, C.; Schmeder, A.; Liu, F.; Wissink, K.; et al. 2026. Systematic benchmarking demonstrates large language models have not reached the diagnostic accuracy of traditional rare-disease decision support tools. *European Journal of Human Genetics*, 34(4): 498–504.

Robinson, P. N.; Ravanmehr, V.; Jacobsen, J. O.; Danis, D.; Zhang, X. A.; Carmody, L. C.; Gargano, M. A.; Thaxton, C. L.; Karlebach, G.; Reese, J.; et al. 2020. Interpretable clinical genomics with a likelihood ratio paradigm. *The American Journal of Human Genetics*, 107(3): 403–417.

Shyr, C.; Cassini, T. A.; Tinker, R. J.; Byram, K. W.; Embí, P. J.; Bastarache, L.; Peterson, J. F.; Xu, H.; and Hamid, R. 2025. Large language models for rare disease diagnosis at the undiagnosed diseases network. *JAMA network open*, 8(8): e2528538.

Smedley, D.; Jacobsen, J. O.; Jäger, M.; Köhler, S.; Holtgrewe, M.; Schubach, M.; Siragusa, E.; Zemojtel, T.; Buske, O. J.; Washington, N. L.; et al. 2015. Next-generation diagnostics and disease-gene discovery with the Exomiser. *Nature protocols*, 10(12): 2004–2015.

Vasilevsky, N. A.; Toro, S.; Matentzoglu, N.; Flack, J. E.; Mullen, K. R.; Hegde, H.; Gehrke, S.; Whetzel, P. L.; Shwetar, Y.; Harris, N. L.; et al. 2026. Mondo: integrating disease terminology across communities. *Genetics*, 232(4): iyaf215.

Wang, H.; Zhao, Y.; Zheng, Y.; and Wu, X. 2025. RareSyn: Health Record Synthesis for Rare Disease Diagnosis. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 12322–12338.

Yao, S.; Shinn, N.; Razavi, P.; and Narasimhan, K. 2024. t-bench: A Benchmark for Tool-Agent-User Interaction in Real-World Domains. arXiv:2406.12045.

Zhao, M.; Havrilla, J. M.; Fang, L.; Chen, Y.; Peng, J.; Liu, C.; Wu, C.; Sarmady, M.; Botas, P.; Isla, J.; et al. 2020. Phen2Gene: rapid phenotype-driven gene prioritization for rare diseases. *NAR genomics and Bioinformatics*, 2(2): lqaa032.

Zhao, W.; Wu, C.; Fan, Y.; Qiu, P.; Zhang, X.; Sun, Y.; Zhou, X.; Zhang, S.; Peng, Y.; Wang, Y.; et al. 2026. An agentic system for rare disease diagnosis with traceable reasoning. *Nature*, 651(8106): 775–784.