When Does Retrieval Beat Direct LLM Diagnosis in Rare Disease? An Empirical Study of Ontology Coverage

Mohamed Elmofty
Humboldt-Universität zu Berlin
mohamed.elmofty@hu-berlin.de

Ulf Leser
Humboldt-Universität zu Berlin
ulf.leser@hu-berlin.de

Abstract

Recent high-complexity agentic systems such as DeepRare perform strongly on rare disease diagnosis benchmarks, but it remains unclear when gains come from structured knowledge access and when they come from parametric LLM knowledge. We compare phenotype-based retrieval, LLM reranking, and unrestricted LLM diagnosis across seven benchmarks covering 10,382 cases. We find a clear performance crossover driven by retrieval coverage—the fraction of cases whose true diagnosis is within the retriever's top-50: on high-coverage datasets, ontology-based retrieval dominates; on low-coverage datasets, open-ended LLM diagnosis takes the lead. Building on this, adding an LLM reranker over retrieved candidates further improves accuracy across our patient-case benchmarks, closing most of the remaining gap to agentic systems (within 2 pp on MME and LIRICAL). We trace the crossover to two structural failure modes of ontology-based retrieval—annotation sparsity and phenotypic homogeneity—and show that aggregate scores across mixed benchmarks can hide these qualitatively different diagnostic settings. These findings motivate per-dataset evaluation and hybrid diagnostic systems that combine retrieval, reranking, and parametric LLM generation based on case characteristics.

1 Introduction

Rare diseases collectively affect over 300 million people worldwide (Wakap et al., 2020), yet patients frequently endure a "diagnostic odyssey" exceeding five years (Schieppati et al., 2008). The sheer number of rare diseases—over 10,000 with known genetic etiology (Haendel et al., 2020)—and limited clinician familiarity with any individual condition make automated diagnostic support a pressing need.

Phenotype-driven diagnosis tools address this by matching patient symptoms, encoded as Human Phenotype Ontology (HPO) terms (Köhler et al., 2021), against curated disease–phenotype annotations. PhenoDP (Wen et al., 2025), a recent deep-learning-based toolkit, combines information-content (IC)-based, phi-coefficient-based, and semantic similarity measures over the HPO Annotation (HPOA) knowledge base to rank candidate diseases, achieving state-of-the-art results among phenotype-based methods.

At the other end of the complexity spectrum, the DeepRare system (Zhao et al., 2026) represents a multi-agent architecture integrating over 40 specialized tools, more than 10 biomedical knowledge sources, web-scale retrieval, and self-reflective reasoning to achieve 57.2% average Recall@1 across seven rare-disease benchmarks. This raises a natural question: *how much of this performance comes from querying curated disease–phenotype ontologies versus relying on knowledge encoded in model parameters?*

We investigate this through two complementary experiments. First, we evaluate a simple two-stage pipeline—PhenoDP retrieval followed by LLM reranking—and compare it against both unrestricted LLM diagnosis and DeepRare's per-dataset results. On the high-coverage patient-case benchmarks, the reranking pipeline is competitive with DeepRare (70.0% vs. 70.0% on MME; 54.1% vs. 56.0% on LIRICAL), despite using only two components and no runtime knowledge search beyond PhenoDP's pre-indexed HPOA annotations. On the DDD (Deciphering Developmental Disorders) disease-profile benchmark, retrieval also scores higher than DeepRare's reported number (59.0% vs. 43.0%), but DDD consists of disease-level HPO bundles rather than patient cases and overlaps ontologically with HPOA, so we treat this more as an upper bound on how well ontology-aligned benchmarks can be matched than as a head-to-head capability comparison (§3.5).

Second, comparing the two diagnostic approaches—ontology-based retrieval and unre-

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 508–518

stricted LLM generation—across seven datasets reveals a *performance crossover*: which approach wins is determined by *retrieval coverage*, the fraction of patient presentations for which PhenoDP successfully places the true diagnosis in its top-50 candidates. Coverage reflects whether the retriever can successfully match a patient's HPO terms to the correct disease; it fails primarily when the target disease is sparsely characterized in the HPOA knowledge base, and secondarily when the patient's phenotype set is small or when the target disease is phenotypically indistinguishable from related conditions. On high-coverage datasets, ontology-based retrieval dominates; on low-coverage datasets, the unrestricted LLM takes the lead. This crossover is invisible in aggregate metrics.

This paper makes four contributions. First, we show that LLM reranking improves on pure ontology-based retrieval on every patient-case benchmark we evaluate; on the high-coverage benchmarks MME (70.0% vs. 70.0%) and LIRICAL (54.1% vs. 56.0%), an open-weight two-component pipeline reaches numbers comparable to DeepRare, although the small sample sizes (n=40 for MME) and unmatched evaluation conditions mean these comparisons should be read as approximate rather than head-to-head (§3.5, Limitations). Second, we observe a performance crossover between ontology-based retrieval and unrestricted LLM diagnosis that tracks retrieval coverage: in our seven benchmarks, datasets above 70% coverage favor retrieval and datasets below 60% favor unrestricted LLM generation. Because no benchmark in our study falls in the 60–70% transition zone, the boundary should be read as a hypothesis consistent with these observations rather than a located threshold. Per-dataset comparisons are supported by McNemar tests and bootstrap CIs. Third, through error analysis of non-retrievable cases, we identify annotation sparsity and phenotypic homogeneity as the two structural failure modes that determine where ontology-based retrieval breaks down. Finally, an oracle analysis shows that the two approaches succeed on largely different cases, and an empirical evaluation of a *pooled* score-based threshold router shows that a single retriever-confidence threshold pooled across datasets fails to close the oracle gap—pointing to input-modality features, or to per-dataset thresholding which we did not evaluate, as candidate routing signals.

2 Related Work

Automated diagnosis for rare diseases has been pursued along two tracks: phenotype-based retrieval tools that operate over curated ontologies, and more recently, general-purpose LLMs applied to clinical reasoning. Our work connects these two lines.

On the retrieval side, tools such as Exomiser (Smedley et al., 2015), LIRICAL (Robinson et al., 2020), and PubCaseFinder (Fujiwara et al., 2018) compute semantic similarity between a patient's HPO terms and the annotations of candidate diseases in the HPOA knowledge base (Köhler et al., 2021); PubCaseFinder additionally indexes published case reports to widen its evidence base. PhenoDP (Wen et al., 2025), which we use as the retrieval backbone in this work, continues this line by combining information-content-based, phi-coefficient-based, and learned semantic similarity. Ontology-anchored methods of this kind are bounded by the completeness of HPOA annotations, so diseases or cases that are under-annotated in HPOA or phenotypically similar to many others are systematically hard to retrieve.

Retrieve-and-rerank pipelines are a standard pattern in open-domain QA and information retrieval, where a fast recall-oriented retriever is followed by a stronger but more expensive reranker (Karpukhin et al., 2020; Nogueira and Cho, 2019). We apply this pattern to ontology-based rare disease diagnosis.

On the LLM side, general-purpose models have been evaluated for clinical reasoning both closed-book (Singhal et al., 2023; Nori et al., 2023) and in rare-disease-specific settings (Chen et al., 2024; Shyr et al., 2024). Domain-adapted biomedical LMs (Luo et al., 2022; Christophe et al., 2024; Google DeepMind, 2025) extend this line. A separate thread couples LLMs with retrieval: retrieval-augmented generation is increasingly used in biomedical question answering (Xiong et al., 2024), typically grounding LLMs in unstructured literature rather than curated ontologies. DeepRare (Zhao et al., 2026) combines LLM reasoning with more than 40 external tools and over 10 biomedical knowledge sources, achieving state-of-the-art performance on several rare-disease benchmarks.

Most closely related to our work are systems that combine an HPO-based phenotype retriever with downstream refinement, but to our knowledge no prior work has systematically compared retriever-only, retrieve-and-rerank, and unrestricted LLM di-

agnosis across a comparable range of benchmarks, or characterized *when* each approach has the advantage and why.

3 Methods

3.1 Task and Metrics

Given a clinical vignette with HPO-encoded phenotypes rendered as natural language, produce a ranked list of OMIM disease diagnoses. We report Recall@k (R@k) and MRR under **strict OMIM ID matching**. Statistical significance: paired McNemar's test; uncertainty: bootstrap 95% CIs (10,000 resamples). Code, prompts, and per-case outputs are released at https://github.com/mofty8/BioNLP26.

We define **coverage** as the fraction of cases where the truth disease appears in PhenoDP's top-50 retrieved candidates. This is a property of each case—a case is *covered* if PhenoDP ranks its true diagnosis within the top 50, regardless of how many HPO terms the patient presents with. Cases where truth is absent from the top 50 are **non-retrievable**—they represent a hard ceiling for any retrieval-based approach that uses the same candidate pool. We measure coverage using PhenoDP as a concrete retriever, but the failure modes we document (§5.1) are properties of HPO/HPOA-anchored retrieval in general and are not specific to PhenoDP's scoring function.

3.2 Retriever: PhenoDP

PhenoDP (Wen et al., 2025) is a deep-learning-based toolkit whose Ranker module scores candidate diseases by combining three similarity measures between a patient's HPO terms and each disease's HPOA profile: information-content (IC) similarity, phi-coefficient similarity after graph propagation, and semantic similarity via contrastive embeddings. Wen et al. (2025) report that IC dominates, phi contributes, and semantic similarity is minimal; in practice PhenoDP's ranking is driven primarily by term-level information content. The Ranker searches the full HPOA knowledge base (12,717 annotated diseases) and returns up to 50 candidates per case; its performance is therefore bounded by HPOA completeness—a limitation central to our analysis.

3.3 Reranking and Unrestricted Diagnosis

We evaluate the LLM under two conditions that share an input format but differ in whether a retriever is present.

Retrieve-and-rerank. The LLM reranker receives (i) a clinical narrative generated by PP2Prompt (Reuter et al., 2024), which converts HPO terms into natural language, and (ii) PhenoDP's top-k candidates ($k \in \{3, 5, 10\}$). Unless otherwise noted, we report top-10 results as this consistently yielded the best performance. We evaluate two prompt strategies: a *rule-gated* prompt with explicit conditional rules (preserve the retriever's ordering when the top-1 score exceeds rank-2 by a large margin, unless specific phenotypic contradictions are identified; full template in Appendix D), and a *narrative* prompt adapted from PP2Prompt (Reuter et al., 2024) that presents candidates without conditional gates and allows free reordering.

Unrestricted LLM diagnosis. The LLM receives only the clinical narrative and produces up to 10 ranked diagnoses from parametric knowledge, with no retriever input.

3.4 Models

For reranking: Gemma-3 27B (Gemma Team, 2025), Gemma-4 31B,[1] Llama-3.3 70B (Meta AI, 2024). For unrestricted diagnosis: additionally Med42-70B and Med42-8B (Christophe et al., 2024), MedGemma-27B (Google DeepMind, 2025), and Qwen2.5-32B (Qwen Team, 2024); Table 2 shows the two strongest open-LLM baselines, and full results for all seven models are in Appendix A. All open-weight, instruction-tuned, temperature 0.

3.5 Datasets

We evaluate on seven benchmarks (Table 1): Phenopackets (Wen et al., 2025), four RareBench subsets (Chen et al., 2024) (LIRICAL, Matchmaker Exchange [MME], RAMEDIS, Hereditary Monogenic Syndromes [HMS]), MyGene2 (Chong et al., 2020), and DDD[2]—curated disease-level HPO profiles from the Deciphering Developmental Disorders database, the same source used by DeepRare. Our DDD subset (n=2,248) differs slightly

[1] Gemma-4 (https://blog.google/innovation-and-ai/technology/developers-tools/gemma-4/); no separate technical report was available at the time of this evaluation, so we cite the official announcement and release the exact checkpoint identifier alongside our code.

[2] https://www.deciphergenomics.org/ddd/ddgenes

Dataset	Type	n	Dis.	Cov (%)	HPOA med
Phenopackets	Lit	6901	500	81.5	47
LIRICAL	Lit	370	411	85.7	43
MME	Clin	40	28	95.0	47
DDD	Curated	2248	2171	79.2	22
MyGene2	CR	131	37	70.2	42
RAMEDIS	CR†	624	160	56.7	31
HMS	Clin	68	101	48.5	50

Table 1: Datasets (10,382 cases total). Shaded rows are *low-coverage* (<60%). **Type** codes: Lit = published case reports with author-provided HPO terms; Clin = prospectively collected hospital records; CR = case reports with HPO terms mapped post hoc; Curated = disease-level HPO profiles from DECIPHER (DDD is not directly comparable to patient-level benchmarks; see §3.5). **Columns**: n evaluable cases; **Dis.** unique truth diseases; **Cov** retrieval coverage (fraction of cases whose truth is in PhenoDP's top-50); **HPOA med** median HPOA annotations for the truth diseases. †RAMEDIS HPO terms were manually mapped from clinical descriptions by Chen et al. (2024). HMS uses the n=68 OMIM-evaluable subset; MyGene2 uses the n=131 PhenoDP-compatible subset.

from DeepRare's (n=2,283) due to filtering for PhenoDP-compatible OMIM diseases; we were unable to exactly reproduce their case selection as the original split is not publicly available. For HMS, the original dataset contains 88 cases; 20 are annotated with ORPHA-only IDs corresponding to non-Mendelian immune disease groupings that do not map to specific OMIM entries and therefore lack the HPOA phenotype annotations required by the retrieval stack, leaving 68 evaluable cases.

4 Results

4.1 Cross-Approach Comparison

Table 2 presents results across the three approaches (retriever alone, retrieve-and-rerank, unrestricted LLM) alongside DeepRare's published numbers. DeepRare results are taken directly from their supplementary table (Zhao et al., 2026); direct numerical comparison is approximate throughout, as DeepRare does not report Phenopackets results and evaluation conditions may differ from ours. Two main patterns emerge.

The reranker improves over retrieval. On Phenopackets, the best reranker (Gemma-4 31B, rule-gated, top-10) improves R@1 from 51.2% to 54.3% (+3.1 pp, McNemar $p < 0.0001$; 262 promotions vs. 67 demotions). On MME, +5.0 pp (65.0% → 70.0%). On LIRICAL, +1.4 pp. On RAMEDIS, the narrative prompt achieves the largest gain: +12.9 pp (10.3% → 23.2%, Gemma-3 27B). The reranker does not improve on DDD (59.0% → 59.0%): DDD is a curated disease-level benchmark whose phenotype profiles are drawn from the same ontological sources as HPOA, so the ranking the retriever produces largely reflects alignment between two versions of the same knowledge base (see §3.5 and Limitations). The combined pipeline reaches numbers comparable to DeepRare DS-V3 on MME (70.0% vs. 70.0%, n=40) and is within 2.5 pp of DeepRare GPT-4o (72.5%), and is similarly close on LIRICAL (54.1% vs. 56.0%). We caution that these comparisons are approximate rather than head-to-head: our pipeline uses open-weight models up to 31–70B against DeepRare's GPT-4o / DeepSeek-V3 backbones, DeepRare's case selection and prompting may differ from ours, and the MME and LIRICAL sample sizes (40 and 370) give wide bootstrap CIs (Table 3). The point is that an open-weight two-component pipeline is in the same neighborhood on high-coverage datasets, not that it is statistically equivalent.

The performance crossover across datasets. On MME, retrieval achieves 65.0% versus the unrestricted LLM's 5.0% (a 60.0 pp gap in favor of retrieval). On RAMEDIS, the LLM achieves 46.8% versus the retriever's 10.3% (a 36.5 pp gap in the opposite direction). This crossover is statistically significant on six of seven benchmarks (McNemar $p < 0.0001$; Table 3); HMS (n=68 shared cases) does not reach conventional significance (p=0.052) due to limited statistical power.

4.2 The Regime Shift

Figure 1 visualizes the regime shift from two angles. The scatter plot (a) shows retrieval coverage is strongly associated with which approach dominates, with the datasets in our study separating into two clusters: five at ≥70% coverage and two at ≤57%, with no observations in the 60–70% range. The boundary should therefore be read as a hypothesis consistent with these data rather than a located threshold; identifying its precise location requires benchmarks in the transition zone. The bar chart (b), with datasets ordered by decreasing coverage, makes the crossover visible: reading left to right, the dominant system switches from retrieval (blue) to unrestricted LLM (orange).

		Pheno.		LIRICAL		MME		DDD		MyG2		RAMED.		HMS	
System	**Model**	**R@1**	**R@5**	**R@1**	**R@5**	**R@1**	**R@5**	**R@1**	**R@5**	**R@1**	**R@5**	**R@1**	**R@5**	**R@1**	**R@5**
Retriever	PhenoDP	51.2	67.2	52.7	70.8	65.0	85.0	59.0	70.2	38.9	61.8	10.3	28.7	7.4	25.0
Reranker	Gem-4 31B	**54.3**	67.6	54.1	75.9	70.0	90.0	59.0	70.2	48.9	61.8	18.3	37.0	14.7	36.8
Reranker	Lla-3.3 70B	52.6	67.3	54.1	75.9	65.0	90.0	**59.1**	70.2	38.9	61.8	22.9	37.0	16.2	36.8
Open LLM	Lla-3.3 70B	14.3	25.6	17.0	28.4	5.0	10.0	12.6	23.1	3.8	15.3	46.0	66.7	17.6	30.9
Open LLM	Med42 70B	11.5	24.1	15.9	26.2	2.5	2.5	6.4	12.8	2.3	16.0	46.8	64.1	20.6	27.9
DeepRare	DS-V3	—[a]	—	56.0	—	70.0	—	43.0	—	72.6	—	72.6	—	**57.0**	—
DeepRare	GPT-4o	—	—	**56.0**	—	**72.5**	—	43.0	—	**75.9**	—	**73.2**	—	54.5	—

Table 2: Cross-approach comparison. Shaded rows: our pipeline. **Bold**: best R@1 per dataset across all systems. The reranker rows report the best (prompt, top-k) configuration on each dataset's test set (see Table 4 and Limitations); fixed-configuration sensitivity is reported for Phenopackets in Appendix B. Green rows: DeepRare from Zhao et al. (2026); DeepRare results are not available for Phenopackets ([a]) and direct comparison is approximate for all datasets (§3.5). All our metrics use strict OMIM ID matching. All HMS numbers use the n=68 OMIM-evaluable subset; all MyGene2 numbers use the n=131 PhenoDP-compatible subset (§3.5). (—): R@5 not available for DeepRare.

Dataset	**Ret**	**LLM**	**CI$_{\text{Ret}}$**	p	**Setting**
Pheno.	51.2	11.5	50.0–52.4	<.001	Retrieval
LIRICAL	52.7	15.9	47.6–57.8	<.001	Retrieval
MME	65.0	2.5	50.0–80.0	<.001	Retrieval
DDD	59.0	12.6	57.7–60.3	<.001	Retrieval
MyGene2	38.9	3.8	30.7–47.5	<.001	Retrieval
RAMEDIS	10.3	46.8	8.0–12.7	<.001	LLM
HMS	7.4	20.6	1.5–14.7	.052	—[†]

Table 3: Statistical significance of the retrieval vs. unrestricted LLM comparison (McNemar test on cases shared between both systems; LLM = Med42 70B). CI: bootstrap 95% for retriever R@1. MyGene2 numbers use the n=131 PhenoDP-compatible subset. [†]HMS: n=68, underpowered.

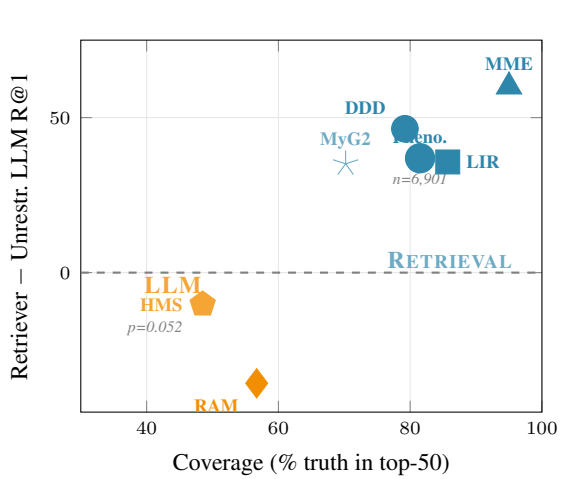
Dataset	**Config**	**Ret**	**Rer**	**Δ**	**P**	**D**
Retrieval regime						
Pheno.	G4/Rule/10	51.2	**54.3**	+3.1	262	67
LIRICAL	G4/Narr/10	52.7	**54.1**	+1.4	10	5
MME	G4/Narr/10	65.0	**70.0**	+5.0	4	2
MyGene2	G4/Narr/10	38.9	**48.9**	+9.9	13	0
LLM regime						
RAMEDIS	G3/Narr/10	10.3	**23.2**	+12.9	78	25
HMS	G3/Narr/10	7.4	**16.2**	+8.8	7	0

Table 4: Reranker performance by dataset, reporting the *best (model, prompt, top-k) configuration on each dataset's test set*; these numbers therefore upper-bound what a single fixed configuration would obtain and should not be read as a held-out estimate (see Limitations and Appendix B for the full Phenopackets grid). **Config** format: Model/Prompt/top-k, where G4=Gemma-4 31B, G3=Gemma-3 27B, L3=Llama-3.3 70B; Rule=rule-gated prompt, Narr=narrative prompt; 10=top-10 candidates. **Ret/Rer**: retriever/reranker R@1 (%). **P**: promotions (reranker moves truth to a higher rank). **D**: demotions (reranker moves truth to a lower rank). DDD is excluded: all configurations yield Δ=0.0 pp. The narrative prompt dominates in the LLM-dominant setting; the rule-gated prompt is optimal in the retrieval-dominant setting (Phenopackets).

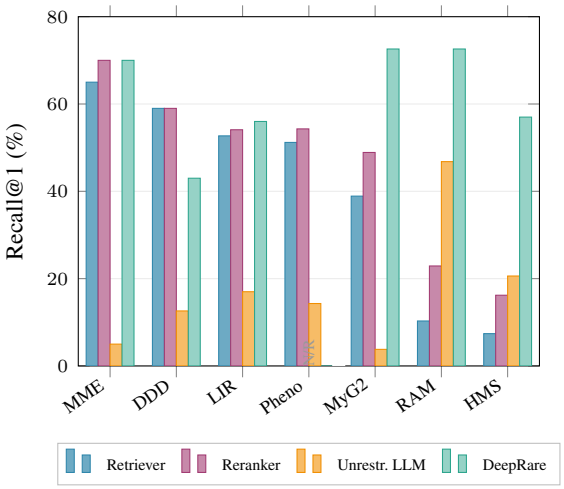
4.3 Reranker Analysis: Per-Subset and Conditional

Table 4 disaggregates the reranker analysis by dataset (DDD excluded, as discussed in §4). First, the reranker improves R@1 on every dataset, with gains ranging from +1.4 pp (LIRICAL) to +12.9 pp (RAMEDIS). On Phenopackets, the 3.9:1 promotion-to-demotion ratio is driven by clean cases like neurofibromatosis type 1 (204 in-window cases: 103 promotions to rank 1, zero demotions); the most-harmed disease is Aarskog-Scott syndrome (15 promotions against 48 demotions), which is consistently confused with Noonan syndrome—the two share facial dysmorphism and short stature but have distinct causal genes. Second, prompt strategy interacts with the dominant approach: the rule-gated prompt excels on Phenopackets (where retrieval is strong and conservative reranking avoids harmful demotions), while the narrative prompt is necessary on RAMEDIS, HMS, and MyGene2 (where the retriever has more room for improvement). Third, the rule-gated prompt is the strongest configuration only on Phenopackets, where retriever scores are well-separated and its conditional rules are triggered; on the shorter RareBench narratives the narrative prompt is chosen as best in Table 4.

Conditional performance on retrievable cases. The overall R@1 numbers in Table 2 mix two effects: how often the retriever places the correct answer in its top-k window (coverage), and how well



(a) Coverage is strongly associated with the dominant approach. All datasets with >70% coverage fall in the retrieval-dominant setting; both with <60% fall in the LLM-dominant setting. Unrestricted LLM: Llama-3.3 70B.



(b) R@1 by approach (decreasing coverage left→right). The crossover is visible. DeepRare adds most value in the LLM-dominant setting (right). Phenopackets not reported by DeepRare.

Figure 1: The diagnostic regime shift. (a) In our seven benchmarks, coverage is associated with which approach dominates; observations cluster at ≥70% or ≤57%, with no data in the 60–70% transition zone, so the crossover boundary is a hypothesis consistent with the data rather than a located threshold. (b) Grouped bars reveal the crossover between approaches.

the reranker reorders candidates once the correct answer is present. To isolate the reranker's discrimination ability, we report a *conditional* ("fair") evaluation restricted to cases where the truth disease is within the retriever's top-k—i.e., cases the retriever could in principle solve and that the reranker actually sees. On RareBench (Llama-3.3, narrative, top-10; LIRICAL + MME + RAMEDIS + HMS, 1,102 cases total), the reranker achieves 65.4% R@1 on the 573 cases whose truth is within the top-10 window, versus the retriever's 50.6% on the same subset (+14.8 pp). On Phenopackets (Gemma-4, rule-gated, top-10), 75.5% versus 71.2% (+4.3 pp). These conditional gains are substantially larger than the overall gains in Table 2, confirming that the LLM adds real clinical reasoning value when the correct answer is already in the candidate set—the modest overall gains are partly an artifact of the hard retrieval ceiling imposed by non-retrievable cases.

5 Explaining the Regime Shift

5.1 Non-Retrievable Cases

The regime shift is directly linked to *non-retrievability*. Table 5 shows the mapping is sharp: the two LLM-regime datasets (RAMEDIS 43.3%, HMS 51.5%) have non-retrievable rates clearly above the retrieval-regime range (5.0–29.8%), and the ordering by non-retrievable rate matches the ordering by which approach wins.

Dataset	Total	Non-ret	Rate	Regime
MME	40	2	5.0%	Retrieval
LIRICAL	370	53	14.3%	Retrieval
Phenopackets	6,901	1,278	18.5%	Retrieval
MyGene2	131	39	29.8%	Retrieval
RAMEDIS	624	270	43.3%	LLM
HMS	68	35	51.5%	LLM

Table 5: Non-retrievable rates align with the regime. DDD is excluded as a curated disease-level benchmark rather than a patient-case dataset (§3.5).

Analysis of the 1,278 non-retrievable cases in the Phenopackets benchmark reveals two structural failure modes:

Annotation sparsity. Non-retrievable diseases have median 24 HPOA entries versus 52 for retrievable ones. Diseases with <20 annotations have 22–37% non-retrievable rates; those with >40 are retrieved at 88.4%. Example: *congenital adrenal hyperplasia (21-hydroxylase deficiency)*—50 cases, only 12 HPOA annotations, 100% non-retrievable.

Phenotypic homogeneity. Developmental and epileptic encephalopathy (DEE) subtypes share near-identical HPO profiles despite distinct genetics: DEE-4 (430 cases, 48% non-retrievable), DEE-11 (260 cases, 82%). HPO-based retrieval cannot disambiguate them. Homogeneity failures are dis-

Dataset	Ret	LLM	Oracle‡	Δ	LLM-only†
Pheno.	51.2	11.5	**56.7**	+5.5	376
LIRICAL	52.7	15.9	**59.7**	+7.0	26
MME	65.0	2.5	**67.5**	+2.5	1
DDD	59.0	12.6	**61.8**	+2.8	63
MyGene2	38.9	3.8	**41.2**	+2.3	3
RAMEDIS	10.3	46.8	**51.1**	+4.3	255
HMS	7.4	20.6	**26.5**	+5.9	13

Table 6: Complementarity of retrieval and unrestricted LLM (Med42 70B) across all seven benchmarks. ‡**Oracle**: R@1 of a hypothetical perfect router that always selects whichever approach answers each case correctly—an upper bound on hybrid system performance. Δ: gain over the better single system. †**LLM-only**: cases answered correctly by the unrestricted LLM but not by the retriever—i.e., cases where only parametric LLM knowledge suffices. Shaded rows: LLM-dominant datasets. Complementarity holds on every dataset; the largest per-dataset oracle gains are LIRICAL +7.0, Phenopackets +5.5, and HMS +5.9.

sociable from sparsity: Xia-Gibbs syndrome is 86% non-retrievable despite 187 HPOA annotations (Appendix C), because its dense profile overlaps with many other syndromic neurodevelopmental disorders.

At category level, epilepsy/neurological (40.2%) and immunodeficiency (42.0%) are highest risk, driven by sparse annotations and phenotypic overlap. Mitochondrial diseases (1.9%) and RASopathies (1.3%) are rarely missed, owing to distinctive phenotypes and rich annotations (median 113 and 56 HPOA entries, respectively). The same sparse-annotation signature appears on other benchmarks: on HMS, non-retrievable diseases have a mean of 8.0 HPOA entries vs. 23.0 for retrievable ones; on RAMEDIS, 23.9 vs. 36.2.

5.2 Complementarity of Retrieval and LLM Diagnosis

Table 6 quantifies how much could be gained by a hybrid system that combines both approaches. A hypothetical perfect router—one that selects whichever approach answers each case correctly—would improve over the better single system by 2.3–7.0 pp across all seven datasets. The key insight is that the two approaches succeed on largely non-overlapping cases. On Phenopackets, 376 patient cases (5.5%) are answered correctly only by the unrestricted LLM; these are predominantly non-retrievable cases where the true disease is absent from PhenoDP's top-50 and only parametric LLM knowledge can recover the answer. On RAMEDIS, 255 cases (40.9%) fall in this category, consistent with its high non-retrievable rate.

These results motivate hybrid systems that combine retrieval and LLM generation. We evaluate a score-based router in §6, finding that retriever confidence alone is insufficient to close the gap and that input-modality features are needed.

6 Discussion

Where agentic complexity adds value. On low-coverage datasets, DeepRare substantially outperforms our pipeline (41–50 pp on RAMEDIS and HMS), suggesting that richer runtime knowledge access matters most in these settings. On high-coverage patient-case datasets (MME and LIRICAL, with n=40 and n=370), our open-weight two-component pipeline lands in the same neighborhood as DeepRare (DDD is excluded from this comparison; see §3.5). We stop short of calling the pipeline "competitive" in a statistical sense: at these sample sizes, a single case flip on MME shifts R@1 by 2.5 pp, and our pipeline uses 31–70B open-weight rerankers against DeepRare's GPT-4o / DeepSeek-V3 backbones under evaluation conditions we could not exactly reproduce. The interpretable claim is structural: on high-coverage settings the simpler pipeline reaches a similar operating point with a smaller open-weight backbone, while on low-coverage settings agentic complexity is doing real work that retrieval-plus-reranking cannot replace. This still supports a practical design principle—route high-coverage cases through a lightweight pipeline and reserve the more resource-intensive agentic system for the cases that genuinely need it—but the boundary of when each is sufficient should be set on larger benchmarks under matched evaluation conditions.

Implications for evaluation. Aggregate metrics across heterogeneous benchmarks obscure the regime shift. Two systems with the same overall average can have very different per-dataset profiles—strong on high-coverage cases but weak on low-coverage ones, or the reverse—and these profiles have very different clinical implications. We recommend per-dataset reporting with explicit coverage characterization, so that readers can understand which cases each system actually handles well.

Toward hybrid systems: a negative result on *pooled* score-based routing. The oracle analysis (Table 6) shows that retrieval and unrestricted LLM

generation succeed on largely non-overlapping cases, with a perfect router improving 2.3–7.0 pp over the better single system across all seven datasets. We tested whether a *single* pooled threshold over PhenoDP confidence signals can serve as a practical routing rule: route to the retriever if a confidence statistic exceeds a threshold τ, otherwise escalate to the LLM. We evaluated four candidate signals: (i) the rank-1 similarity score, (ii) the margin between rank-1 and rank-2 scores, (iii) the coefficient of variation (CV) across the 50 candidate scores—a high CV indicates that one disease scores substantially above the others, while a low CV suggests the candidates are tightly clustered—and (iv) the product of rank-1 similarity and CV. A pooled threshold sweep on rank-1 similarity (N=10,379) finds τ^*=0.705, recovering +1.2 pp (16% of the oracle gap) in-sample; in a leave-one-dataset-out evaluation, the router *hurts* on five of seven held-out datasets (micro-average: 49.7% vs. 50.1% retriever baseline). The other three signals behave similarly: none improve over the single-score baseline in the leave-one-dataset-out setting. This result speaks only to pooled thresholding; per-dataset thresholds, learned routers, and confidence signals derived from sources other than PhenoDP's own scores were not evaluated.

Why a pooled threshold fails. The failure is structural rather than a tuning issue. RAMEDIS, where the LLM leads by 35 pp (46% vs. 10%), needs $\tau \to \infty$ (always escalate to the LLM), while Phenopackets and DDD, where retrieval leads by ≈37 and ≈46 pp respectively, need $\tau \to 0$ (never escalate). A single pooled threshold calibrated on the mix is miscalibrated for both, and no monotone confidence signal over the retriever's own scores can resolve this tension with a pooled rule. This does not rule out per-dataset thresholding or learned routers, which we did not evaluate here.

What a useful router would need. Effective case-level routing requires a signal that reflects *input modality*—whether a case is expressed as structured HPO terms (favoring retrieval) or free-text clinical narrative (favoring parametric LLM knowledge)—rather than retriever confidence alone. Candidate features include HPO annotation density, clinical concept density, and the ratio of specific to general phenotype terms. A learned router combining these with the PhenoDP confidence signals above could in principle exploit the 2–7 pp oracle gains identified in Table 6; pairing such a router with the reranker studied here is a natural next step.

Limitations

Several aspects of this study warrant caution.

1. **Benchmark composition.** Our strongest results are on literature-derived and curated benchmarks (Phenopackets, LIRICAL, MME, DDD), which tend to be the easier diagnostic settings (Zhao et al., 2026); the patterns we report may not transfer to more complex, real-world clinical presentations with noisy phenotypes and incomplete workups.
2. **Model scope.** We evaluate only open-weight models up to 70B parameters; frontier proprietary models used as rerankers could yield different results.
3. **DeepRare comparison.** The comparison uses DeepRare's published per-dataset numbers, which may have been obtained under evaluation conditions different from ours (e.g., different case selection, metrics, or prompting); cross-paper comparisons should be treated as approximate.
4. **DDD is not a patient-case benchmark.** Our DDD subset (n=2,248) does not exactly match DeepRare's (n=2,283) due to filtering for PhenoDP-compatible OMIM diseases. Beyond this case-selection gap, DDD is a curated disease-level HPO profile benchmark rather than a set of individual patient cases, and its phenotype annotations are drawn from the same ontological sources as HPOA, so any retrieval score on DDD partly reflects alignment between two versions of the same knowledge base and overstates the ceiling for real patient cases.
5. **HMS statistical power.** HMS has only 68 evaluable cases (after excluding the 20 ORPHA-only cases without OMIM/HPOA entries), which limits statistical power; our McNemar test on HMS does not reach conventional significance (p=0.052).
6. **Seven-dataset coverage axis.** The coverage-based account of the regime shift is built on seven datasets. Additional benchmarks—particularly ones with moderate coverage in the 60–70% range that lies in the transition zone—would strengthen the empirical case.

7. **Per-dataset configuration selection.** The reranker numbers we report in Tables 2 and 4 use the best (model, prompt, top-k) configuration selected on each dataset's test set. They therefore upper-bound what a single fixed configuration would achieve and do not constitute a held-out estimate. We report a full Phenopackets configuration grid in Appendix B to characterize sensitivity, but analogous grids on the remaining datasets and leave-one-dataset-out configuration selection were beyond the scope of this study.

8. **Retrospective evaluation.** All benchmarks used here have known ground truth. We do not evaluate how any of these systems perform prospectively in clinical workflows, where the distribution of cases and the cost of errors may differ substantially from benchmark conditions.

Ethics Statement

All datasets used here are publicly available and, where applicable, released under protocols covering research use; we did not collect or contact patients, and no new protected health information was generated.

This work is a retrospective methodological study, not a clinical system: none of the pipelines we evaluate are safe for deployment without prospective validation, human oversight, and local calibration. Rare disease diagnosis has asymmetric costs—missed diagnoses prolong the diagnostic odyssey, incorrect ones can trigger unnecessary interventions—so any deployment should treat system output as a suggestion to a clinician, not an answer, and should abstain on low-confidence cases and on regimes outside the system's demonstrated competence (e.g., the low-coverage regime identified here).

Both components inherit biases from their training data. HPOA is not uniformly complete, and the resulting annotation gaps drive the non-retrievability failures we study, producing structural performance disparities across diseases and, by extension, across the patient groups most affected by them. Parametric LLMs over-represent literature-reported presentations, and our benchmarks skew toward published case reports and curated profiles; the numbers here should not be read as predictive of performance on pediatric, low-resource, or non-Western presentations.

Code, prompts, and per-case outputs are released at https://github.com/mofty8/BioNLP26.

Acknowledgments

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – RTG2424/CompCancer – project number 377984878.

References

Xuanzhong Chen, Xiaohao Mao, Qihan Guo, Lun Wang, Shuyang Zhang, and Ting Chen. 2024. RareBench: Can LLMs serve as rare diseases specialists? In *Proceedings of KDD*.

Jessica X Chong and 1 others. 2020. MyGene2: improving the utility of undiagnosed disease web resources. *American Journal of Human Genetics*, 106:287–291.

Clément Christophe and 1 others. 2024. Med42-v2: A suite of clinical LLMs. *arXiv preprint arXiv:2408.06142*.

Toyofumi Fujiwara, Yasunori Yamamoto, Jin-Dong Kim, Orion Buske, and Toshihisa Takagi. 2018. PubCaseFinder: a case-report-based, phenotype-driven differential-diagnosis system for rare diseases. *The American Journal of Human Genetics*, 103:389–399.

Gemma Team. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Google DeepMind. 2025. MedGemma: Medical AI models from Google. *arXiv preprint*.

Melissa Haendel, Nicole Vasilevsky, and 1 others. 2020. How many rare diseases are there? *Nature Reviews Drug Discovery*, 19:77–78.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781.

Sebastian Köhler and 1 others. 2021. The Human Phenotype Ontology in 2021. *Nucleic Acids Research*, 49(D1):D1207–D1217.

Renqian Luo and 1 others. 2022. BioGPT: Generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6).

Meta AI. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage re-ranking with BERT. *arXiv preprint arXiv:1901.04085*.

Harsha Nori and 1 others. 2023. Can generalist foundation models outcompete special-purpose tuning? *arXiv preprint arXiv:2311.16452*.

Qwen Team. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Max S Reuter and 1 others. 2024. Evaluation of LLMs for rare disease differential diagnosis using structured phenotype data. *medRxiv*.

Peter N Robinson, Vida Ravanmehr, Michael A Gargano, and 1 others. 2020. LIRICAL: an approach to prioritize candidate diseases for rare disease diagnostics. *American Journal of Human Genetics*, 107:168–176.

Arrigo Schieppati, Jan-Inge Henter, Erica Daina, and Anita Aperia. 2008. Why rare diseases are an important medical and social issue. *The Lancet*, 371:2039–2041.

Cathy Shyr, Yan Hu, Lisa Bastarache, Alex Cheng, Rizwan Hamid, Paul Harris, and Hua Xu. 2024. Identifying and extracting rare diseases from clinical notes using generative large language models. *Journal of the American Medical Informatics Association*, 31(4):909–918.

Karan Singhal and 1 others. 2023. Large language models encode clinical knowledge. *Nature*, 620:172–180.

Damian Smedley, Julius OB Jacobsen, Maximilian Jäger, and 1 others. 2015. Next-generation diagnostics and disease-gene discovery with the Exomiser. *Nature Protocols*, 10:2004–2015.

Stephanie Nguengang Wakap and 1 others. 2020. Estimating cumulative point prevalence of rare diseases. *European Journal of Human Genetics*, 28:165–173.

Baole Wen, Sheng Shi, Yi Long, Yanan Dang, and Weidong Tian. 2025. PhenoDP: leveraging deep learning for phenotype-based case reporting, disease ranking, and symptom recommendation. *Genome Medicine*, 17:67.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024. Benchmarking retrieval-augmented generation for medicine. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6233–6251.

Weike Zhao, Chaoyi Wu, Yanjie Fan, and 1 others. 2026. An agentic system for rare disease diagnosis with traceable reasoning. *Nature*, 651:775–784.

A Full Unrestricted LLM Comparison

B Full Reranker Grid: Phenopackets

C Non-Retrievable Case Studies

The two structural failure modes identified in §5.1 (annotation sparsity and phenotypic homogeneity) are best understood through concrete examples.

Model	R@1	R@5	R@10	MRR
Llama-3.3 70B	14.3	25.6	29.9	.192
Gemma-4 31B	12.2	19.9	23.1	.156
Med42-70B	11.5	24.1	27.8	.171
Gemma-3 27B	11.3	17.4	20.0	.140
Qwen2.5-32B	10.5	17.1	19.5	.133
MedGemma 27B	10.2	15.8	17.3	.126
Med42-8B	7.1	10.3	11.4	.085

Table 7: Unrestricted LLMs on Phenopackets (strict OMIM ID).

Model	Prompt	K	R@1	Δ	MRR
Gemma-4	Rule	10	**54.3**	+3.1	**.601**
Gemma-4	Rule	5	53.7	+2.5	.590
Gemma-4	Rule	3	53.5	+2.2	.578
Llama-3.3	Rule	10	52.6	+1.4	.589
Llama-3.3	Rule	5	52.3	+1.1	.581
Llama-3.3	Rule	3	52.4	+1.2	.572
Gemma-3	Rule	10	51.6	+0.4	.584
Llama-3.3	Narr	3	51.8	+0.6	.568
Gemma-4	Narr	10	50.7	−0.6	.575
Gemma-3	Narr	10	48.8	−2.5	.565

Table 8: Phenopackets reranker grid (n=6,901, PhenoDP baseline: 51.2%). The rule-gated prompt consistently improves; the narrative prompt can harm on this high-coverage dataset.

Annotation sparsity: congenital adrenal hyperplasia (21-hydroxylase deficiency). This is the single most striking failure in the Phenopackets benchmark: 50 patient cases, all 50 non-retrievable (100%). The disease has only 12 HPO annotations in HPOA—a fraction of the 52-term median for retrievable diseases. Patient phenotype sets typically include electrolyte disturbance, virilization, ambiguous genitalia, salt wasting, and adrenal crisis, but the sparse HPOA profile does not encode enough of these terms for the information-content-based similarity used by PhenoDP to pull the disease into the top-50. An unrestricted LLM, by contrast, recognizes the classic clinical pattern from training text and returns the correct diagnosis immediately. This is a clean illustration of why coverage, not reasoning, is the bottleneck in such cases: no amount of LLM reranking can rescue a candidate that is never retrieved.

Phenotypic homogeneity: the DEE cluster. Developmental and epileptic encephalopathies (DEE-4, DEE-9, DEE-11, and related subtypes) share nearly identical HPO profiles despite distinct underlying genes (KCNQ2, SCN1A, STXBP1, and others). DEE-4 accounts for 430 cases in Phenopackets with a 48% non-retrievable rate; DEE-11 ac-

counts for 260 cases with an 82% non-retrievable rate. When truth is in the top-50, the retriever ranks several phenotypically indistinguishable subtypes together, and the information needed to disambiguate—gene panels, age at onset, EEG pattern, response to specific anti-epileptic drugs—is not encoded in the HPO terms the retriever sees. This is the homogeneity failure mode: retrieval has no mechanism to break ties between entities that look the same in ontology space. Xia-Gibbs syndrome (65 cases, 86% non-retrievable despite 187 HPOA annotations) fails for the same reason at a broader scale—its phenotype profile is dense but overlaps with many other syndromic neurodevelopmental disorders.

Category-level pattern. These case-level failures aggregate into category-level signatures (§5.1). Immunodeficiency (42.0%) and epilepsy/neurological (40.2%) categories are dominated by sparse-annotation or phenotypically overlapping disease families and account for most non-retrievable cases. Mitochondrial diseases (1.9%) and RASopathies (1.3%) are rarely missed—both have distinctive multi-system phenotypes and dense annotations (median 113 and 56 HPOA entries, respectively). The practical implication is that non-retrievability is predictable from disease identity, not random; a routing policy that escalates cases in the high-risk categories to an LLM is therefore a sensible starting point for hybrid systems (§6).

D Prompt Templates

For reproducibility, we include the literal prompt templates used for the rule-gated and narrative reranking conditions. Both prompts consume (i) the PP2Prompt-generated clinical narrative and (ii) PhenoDP's top-k candidates with similarity scores; models are asked to return a ranked list of k OMIM IDs (inference settings are described in §3). The rule-gated prompt enforces three conservative gates (score-gap, obligatory-feature, default-to-retriever) and changes the ranking in $<5\%$ of cases on Phenopackets. The narrative prompt, adapted from PP2Prompt (Reuter et al., 2024), omits the gates and produces 3–5$\times$ more rank changes; it is beneficial on low-coverage datasets but can harm high-coverage ones (e.g., DDD: -4.7 pp).

Rule-gated prompt.

```
You are an expert clinical geneticist.
You will rerank candidate rare diseases
for a patient. Apply the following rules
in order before changing the ranking.

[Score-gap gate]
- If rank-1 score exceeds rank-2 by >= 0.15,
  preserve the retriever's ordering unless
  a Condition A or Condition B violation
  is identified.
- If rank-1 score exceeds rank-2 by >= 0.03
  but < 0.15, you may swap rank-1 and rank-2
  only if a Condition A or Condition B
  violation is identified for rank-1.

[Condition A: obligatory absent feature]
Demote a candidate if ALL four hold:
  (1) the candidate has an obligatory
      HPO feature (frequency "always"
      or "very frequent") in HPOA,
  (2) that feature is in a clinically
      testable category for this patient
      (not age- or assessment-gated),
  (3) the patient record explicitly lists
      the feature as absent (negated),
  (4) absence is not trivially explained
      by the patient's age or workup stage.

[Condition B: biological impossibility]
Demote a candidate if the patient has a
finding that is biologically incompatible
with the candidate disease (e.g., male
patient with an X-linked female-specific
disorder).

[Same-gene spectrum rule]
When multiple candidates differ only in
severity labels for the same gene (e.g.,
"mild" vs. "severe" forms of the same
allelic series), preserve the retriever's
order unless a clear severity marker in
the patient record favors a different form.

[Numbered-subtype rule]
When candidates are numbered subtypes of
the same disease family (e.g., DEE-4 vs.
DEE-11), do not reorder unless a gene,
biomarker, or distinctive phenotype is
explicitly named in the patient record.

Patient narrative:
{narrative}

Candidate diseases (top-{k}):
{candidates_with_scores}

Output: a ranked list of {k} OMIM IDs,
one per line, most likely first. Include
a one-sentence justification for every
change you make relative to the retriever
order, referencing specific HPO terms or
conditions above.
```

Narrative prompt.

```
You are an expert clinical geneticist.
Read the patient narrative carefully and
rank the candidate rare diseases from
most to least likely.

Patient narrative:
{narrative}

Candidate diseases (top-{k}):
{candidates_with_scores}

Produce a ranked list of {k} OMIM IDs,
one per line, most likely first. Base your
ranking on clinical judgment: consider
the full phenotype picture, obligatory
features, and which candidate best
explains the patient. Provide a brief
justification for your top pick.
```

The full prompt strings—including system messages, few-shot examples where used, and the exact candidate-list formatting—are released alongside the code.