RAREDX: CONTROLLED KNOWLEDGE USE AND KNOWLEDGE-GRAPH-GROUNDED POLICY OPTIMIZATION FOR RARE-DISEASE DIAGNOSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Rare-disease diagnosis is a long-tail reasoning problem: phenotypes are incomplete, individual disorders are sparsely documented, and relevant evidence is distributed across ontologies, gene annotations, and biomedical text. Language models consequently favor common conditions, miss rare candidates, or produce plausible but invalid names. We introduce **RareDx**, which couples controlled evidence use with knowledge-graph-grounded policy optimization. RareDx-Harness normalizes heterogeneous records into one ranked-diagnosis task and compares direct inference, static retrieval, adaptive tools, and structured phenotype-gene-disease reasoning over a shared knowledge layer. The training pipeline combines Top-10 post-training with **RareDx-KGPO**, our knowledge-graph-grounded policy optimization method. Its reward projects predictions into a canonical disease graph and integrates curated graded relevance, ontology proximity, biomedical similarity, and phenotype consistency. Vocabulary and output-budget constraints prevent dense partial credit from rewarding fabricated or overlong differentials. Across eight benchmarks, the complete RareDx system centered on Qwen3.5-9B reaches 38.34 macro Hit@10, 1.60 points above GPT-5.5 under the archived protocol; a disjoint validation-selection audit retains a 6.80-point routing gain over Direct on held-out cases. The 27B system reaches 23.53/36.56/40.76 at Hit@1/5/10. Controlled ablations show that retrieval is not uniformly helpful and that controlled routing is central to the gain. These results show that structured medical knowledge can make compact models competitive across heterogeneous long-tail settings. Code and artifacts are available in the anonymous repository.

1 INTRODUCTION

Rare diseases affect fewer than 1 in 2,000 individuals each, yet collectively affect more than 300 million people worldwide Valdez et al. (2016); Nguengang Wakap et al. (2020); Paul (2013); Jonker et al. (2024); Zhao et al. (2026b). Diagnosis commonly takes 4-5 years Ghosh (2025): phenotypes are incomplete, heterogeneous, and shared with common disorders, increasing misdiagnosis risk Dong et al. (2020). Evidence is also dispersed across clinical records, genetic tests, specialist knowledge, and disease resources. A useful system must retrieve the right evidence, reject plausible distractors, and return a ranked differential of recognized disease entities.

Multimodal patient data and curated disease knowledge offer complementary diagnostic evidence Lee et al. (2022). Foundation models, including LLMs Thirunavukarasu et al. (2023) and vision-language models Liu et al. (2025a), further broaden the information accessible to medical AI systems Sarker et al. (2024); Liu et al. (2024); Tran et al. (2025); Du et al. (2025); Liu et al. (2026a; 2025b). RareBench Chen et al. (2024) and RareArena Chen et al. (2026a) evaluate general LLMs, while agentic systems combine retrieval Wang et al. (2025), multi-agent communication Dhatterwal et al. (2023); Chen et al. (2025), and tools. DeepRare Zhao et al. (2026b) and Hygieia Liu et al. (2026b), for example, integrate these components. Structured systems such as AI-MARRVEL Mao et al. (2024), LIRICAL Robinson et al. (2020), and Exomiser Smedley et al. (2015) have demonstrated the value of phenotype, genotype, and disease knowledge for prioritization.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

However, two methodological limitations remain. First, agentic diagnostic systems are often evaluated as monolithic pipelines, making it difficult to determine whether gains arise from the backbone model, retrieval context, tool policy, repeated sampling, or output normalization. Retrieval itself is not uniformly beneficial: irrelevant context can displace decisive evidence even when the knowledge base contains the reference disease. Second, exact-match rewards are too sparse for ranked rare-disease prediction. They assign the same failure signal to a clinically related disorder and an arbitrary error, whereas unconstrained semantic rewards may favor fabricated disease names or indiscriminate candidate enumeration. These limitations are particularly consequential for compact open models, which require reliable domain supervision to approach the diagnostic performance of closed-source LLMs in this long-tail setting reliably.

We introduce RareDx, coupling RareDx-Harness for controlled knowledge use with RareDx-KGPO for knowledge-graph-grounded post-training. The harness compares direct inference, static retrieval, adaptive tools, structured phenotype-gene-disease reasoning, and aggregation under one task and normalizer. RareDx-KGPO projects predictions into a medical graph and rewards graded relevance, ontology proximity, semantic similarity, and phenotype consistency while constraining invalid or overlong outputs. Across eight benchmarks, the complete 9B system reaches 38.34 macro Hit@10, 1.60 points above GPT-5.5 under the archived protocol; a disjoint selection audit retains a 6.80-point gain over Direct. The 27B system exceeds GPT-5.5 at all three cutoffs. Ablations show that retrieval is not uniformly useful and that graph-aware, controlled knowledge use drives the gain.

2 METHODS

2.1 PROBLEM FORMULATION

Given record x, the model returns $\pi(x) = (d_1, \ldots, d_{10})$. The record may contain free text, Human Phenotype Ontology (HPO) terms (Köhler et al., 2021), or genetic findings. For the accepted names and synonyms $\mathcal{G}(x)$, we report the normalized hit indicator

$$\text{Hit@}\, k(x) = \mathbb{1}[\{d_1, \ldots, d_k\} \cap \mathcal{G}(x) \neq \varnothing], \tag{1}$$

for $k \in \{1, 5, 10\}$ after deterministic disease canonicalization. Hit@10 is the primary endpoint because the task requires a differential; Hit@1 isolates first-choice prioritization.

2.2 RAREDX-HARNESS: CONTROLLED DIAGNOSTIC EVALUATION

Figure 1 summarizes RareDx-Harness, which isolates parametric knowledge, retrieval, tool policy, and stochastic aggregation. Every strategy consumes the same normalized case and returns the same Top-10 schema, enabling paired comparison under one judge.

Diagnostic strategies. *Direct* inference isolates the model's parametric knowledge. *Static RAG* retrieves once and prepends a fixed evidence block. *Adaptive ReAct* exposes separate disease and gene search tools, allowing the model to decide what to retrieve and when to stop. *Structured three-hop reasoning* first ranks genes from patient HPO terms, obtains gene-function evidence, and then predicts diseases from the phenotype and gene evidence jointly. For multi-sample inference, reciprocal-rank fusion (RRF) combines lists without requiring calibrated generation probabilities.

Operational router. Routing is dataset-level rather than per patient: each model-benchmark pair chooses between frozen Direct predictions and a deterministic 35B Diagnostic Audit that reranks the same list without external tools. In the leakage-controlled audit, the route is selected on a fixed 20% development partition by Hit@10, then Hit@5 and Hit@1, and evaluated on the remaining 80%. Macro Hit@10 rises from 30.54 to 37.34 for 9B and from 29.83 to 38.13 for 27B. The complete algorithm, prompts, candidate results, and ten-split stability analysis are in Appendix C.

Shared evidence layer. All strategies share 27,554 records integrated from HPO (Köhler et al., 2021), Orphanet (Rath et al., 2012), OMIM (Amberger et al., 2019), and Mondo (Vasilevsky et al., 2026). Dense retrieval uses PubMedBERT, FAISS, and MedCPT (Gu et al., 2021; Douze et al., 2024; Jin et al., 2023); structured reasoning uses the HPO graph, Resnik similarity (Resnik, 1995),

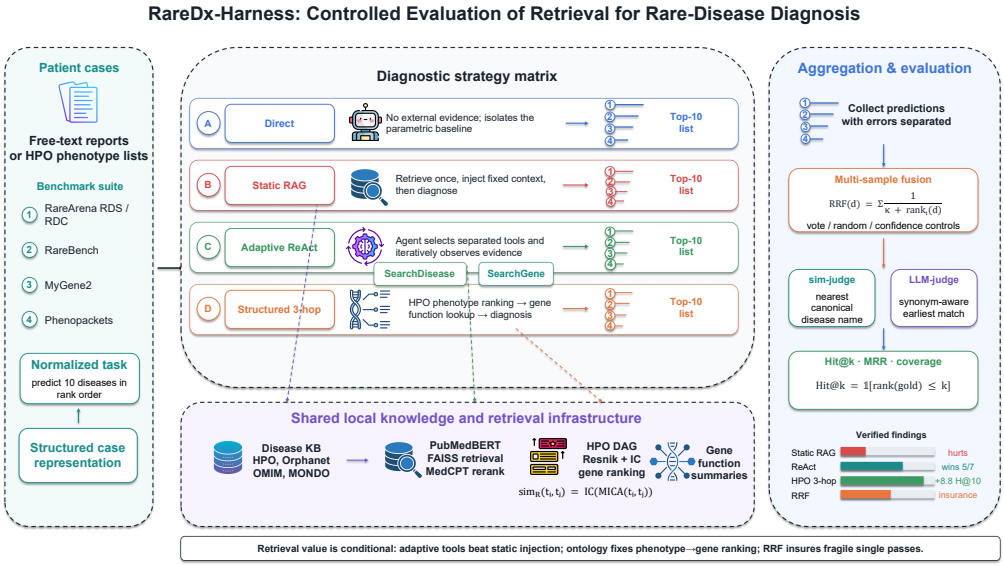

Figure 1: RareDx-Harness. Patient cases from heterogeneous benchmarks are normalized into a common ranked-diagnosis task. The controlled strategy matrix compares direct inference, one-shot static retrieval, adaptive ReAct tool use, and structured phenotype-to-gene-to-disease reasoning over the same local medical knowledge infrastructure. Canonicalization and rank aggregation are shared across strategies to preserve controlled comparison across all branches.

and gene-function summaries. Because evidence and output contracts are shared, performance differences reflect how each strategy uses the same medical knowledge rather than unequal access or incompatible output formatting across methods.

2.3 RareDx-KGPO: knowledge-graph-grounded policy optimization

The two-stage recipe (Figure 2) first teaches a valid Top-10 contract through SFT, then samples multiple lists and computes group-relative advantages from graph-grounded clinical rewards. The contribution is structured medical supervision and its safeguards, not a new policy-gradient estimator. Zero-variance groups are resampled, policy-ratio clips are decoupled (Shao et al., 2024; Yu et al., 2025), and a reference KL term limits distribution shift. Without resampling, uniformly wrong groups supply no relative signal; without asymmetric clipping and reference regularization, dense rewards can shift the policy toward fluent but invalid disease names and away from the diagnostic prior established during supervised fine-tuning.

2.4 Projecting generations into a medical knowledge graph

Each generated string is projected into a canonical space of disease aliases, ontology edges, disease-HPO annotations, HPO ancestors, and information-content statistics. Exact matching is followed by bounded edit-distance snapping; out-of-vocabulary strings retain their rank but receive no dense credit. Every positive reward is therefore traceable to a known disease or supported graph relation.

For candidate y_k at rank k, four complementary evidence channels are evaluated. Curated graded-gold relevance g_k assigns unit credit to the accepted diagnosis and its synonyms and lower credit to verified clinically related diseases. Ontology credit is

$$o_k = 2^{-d_{\text{onto}}(y_k, y^*)}, \tag{2}$$

when a graph path is available. Name-level credit e_k is a thresholded, capped cosine similarity between biomedical entity representations, preventing plausible names from matching exact-diagnosis

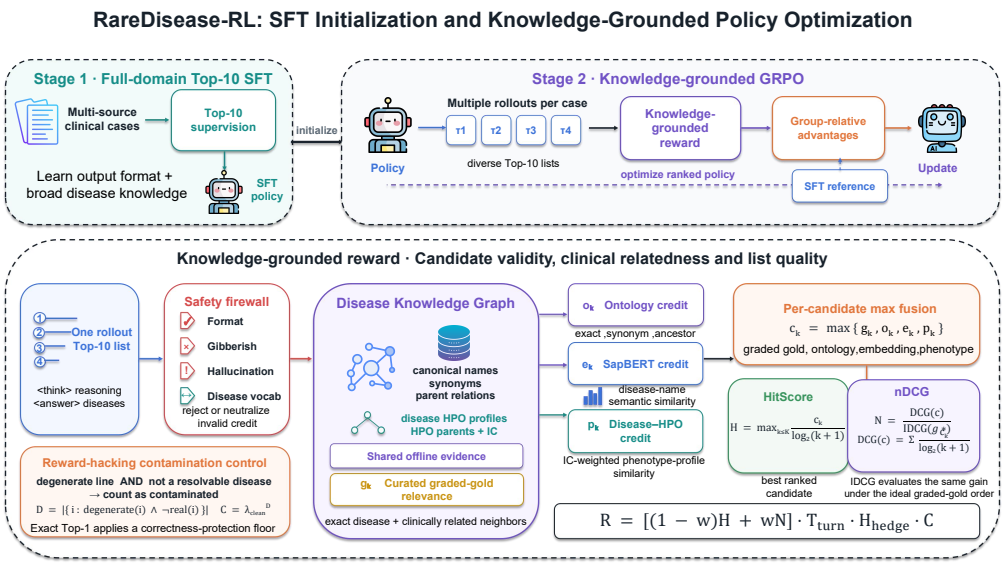

Figure 2: Two-stage post-training. Top-10 SFT establishes the output contract and broad diagnostic knowledge. RareDx-KGPO then optimizes ranked differential diagnoses using disease-graph, semantic, and phenotype evidence, with explicit safeguards against invalid names and reward hacking.

credit. Phenotype credit compares propagated profiles $P(y_k)$ and $P(y^*)$:

$$p_k = \min\left(\tau_p, \frac{\sum_{h \in P(y_k) \cap P(y^*)} \mathrm{IC}(h)}{\sum_{h \in P(y_k) \cup P(y^*)} \mathrm{IC}(h)}\right). \tag{3}$$

Information content emphasizes specific phenotypes; symmetric normalization avoids favoring broad diseases with many nonspecific annotations.

The four channels encode different evidence. Curated labels are precise but sparse, ontology paths are interpretable but incomplete, entity embeddings cover lexical variation but can overvalue plausible phrasing, and phenotype overlap remains informative when disease edges are missing. We therefore use maximum fusion

$$c_k = \max\{g_k, o_k, e_k, p_k\}. \tag{4}$$

This retains the strongest supported relation without double-counting correlated evidence; canonical projection prevents semantic similarity from rescuing fabricated names.

We retain a hit-oriented term

$$H = \max_{k \le K} \frac{c_k}{\log_2(k+1)} \tag{5}$$

and add a list-quality term

$$N = \frac{\mathrm{DCG}\,@K}{\mathrm{IDCG}\,@K} = \frac{\sum_{k=1}^{K} g_k / \log_2(k+1)}{\sum_{k=1}^{K} g_k^* / \log_2(k+1)}, \tag{6}$$

where $g_1^*, \ldots, g_K^*$ is the ideal graded ordering. H prioritizes the best-supported diagnosis, while N separates lists with the same top candidate but different differential quality. Using the same curated scale in its numerator and denominator prevents semantic similarity from inflating nDCG and creates within-case variation. Without graded labels, $w = 0$ recovers H.

For a valid response, the trajectory reward is

$$R_{\mathrm{valid}} = \big[(1-w)H + wN\big] T_{\mathrm{turn}} P_{\mathrm{hedge}} C_{\mathrm{clean}}. \tag{7}$$

Malformed outputs receive zero and globally degenerate outputs receive -1. P_{hedge} penalizes overlong lists and $C_{\mathrm{clean}} = \rho^m$ penalizes m degenerate, unresolvable lines, while valid long disease names and exact Top-1 answers are protected. The single-turn experiments set $T_{\mathrm{turn}} = 1$. Thresholds, transformations, and inactive safeguards are specified in Appendix D.

2.5 EVALUATION SETUP

We evaluate all cases from MyGene2 (Alsentzer et al., 2025), RareBench (Chen et al., 2024) (includes RAMEDIS, MME, HMS, and LIRICAL), the RareArena RDS and RDC tracks (Chen et al., 2026a), and Phenopackets (Danis et al., 2025). Every method returns ten diseases scored by the same deterministic normalizer; macro scores weight the eight sources equally. The principal model is Qwen3.5-9B initialized by Top-10 SFT. Its archived run uses Equation 7 with $w = 0$; the nDCG extension is evaluated only on fixed held-out lists. Qwen series models (Yang et al., 2025), Claude Opus 4.7 (Anthropic, 2026), GLM-5.2 (GLM-5-Team et al., 2026), GPT-5.5 (OpenAI, 2026), and Occamy-1.0 (Chen et al., 2026b) use the identical prompt, parser, and matcher. No training row shares a patient identifier or normalized case text with evaluation; profile-level and seven-benchmark sensitivity audits preserve the main comparison. Dataset sizes, provenance, overlap results, reward thresholds, and split limits appear in Appendices D and E.

3 RESULTS

3.1 OVERALL DIAGNOSTIC PERFORMANCE

Table 1 gives 360 measurements from 15 systems, eight benchmarks, and three cutoffs under one scoring protocol. RareDx-KGPO improves Qwen3.5-9B from 9.53/20.69/24.77 to 16.90/25.54/31.35 at Hit@1/5/10; the harness reaches 20.11/31.54/38.34. It exceeds GPT-5.5 by 1.60 points at Hit@10, is competitive at Hit@5, and remains 1.90 points lower at Hit@1. RareDx-27B reaches 23.53/36.56/40.76 and exceeds GPT-5.5 by 1.52/4.97/4.02 points. These rows use the archived dataset-level route and may invoke the 35B audit. Under disjoint development selection, 9B and Appendix C reports the disjoint-selection result: 9B and 27B reach 37.34/38.13 Hit@10 versus 30.54/29.83 for Direct on matched held-out cases.

Gains concentrate on RAMEDIS and HMS for 9B and additionally on RDS for 27B; MME remains difficult and 9B loses 1.1 Hit@10 points on LIRICAL. This heterogeneity motivates controlled routing. Scale alone is insufficient: direct Qwen3.8-27B reaches 14.4/24.7/28.5, versus 23.5/36.6/40.8 for RareDx-27B. Occamy-1.0 reaches 4.46/11.32/14.65, showing that generic co-work post-training does not directly transfer to this specialization. The three cutoffs also separate candidate coverage from first-choice accuracy. Harness gains grow with k, indicating that controlled evidence primarily improves the clinically useful differential before consistently resolving its top rank. This is preferable to reporting a single cutoff that could conceal shallow prioritization gains or indiscriminate expansion of the differential without improving diagnostic prioritization.

3.2 RL ALGORITHM COMPARISON

GRPO uses sequence rewards and group-relative normalization; DAPO adds dynamic sampling and decoupled clipping; OPD distills dense teacher distributions on student trajectories (Agarwal et al., 2024). RareDx-KGPO instead applies Equation 7 without a rollout teacher; Appendix Table 8 summarizes the supervision and update distinctions.

From the shared initialization in Figure 3, GRPO reaches 0.40 but collapses after step 210 as indiscriminate disease-name generation exploits semantic partial credit. DAPO and OPD remain near 0.40 but plateau or decline. RareDx-KGPO reaches approximately 0.58 without late collapse because vocabulary projection, capped partial credit, graph and phenotype evidence, and anti-hedging penalties close these reward shortcuts. Inspection of late GRPO rollouts confirms the corresponding behavioral change: outputs become weakly conditioned on the patient and enumerate disease names that obtain incidental semantic credit. RareDx-KGPO instead keeps positive credit tied to canonical entities and graph-supported relations, so higher reward remains aligned with a bounded, patient-specific differential rather than a generic disease list.

3.3 REWARD-COMPONENT ABLATION AND ADVERSARIAL AUDIT

On 120 held-out diagnoses, the phenotype channel gives the nearest non-reference HPO neighbor 0.316 mean reward versus 0.008 without it, while exact diagnoses remain at 1.0. The vocabulary gate rejects every constructed pseudo-disease; removing it rewards 82.5%. With rank 1 fixed, the list-

Table 1: Complete diagnostic ranking results (%) on eight benchmarks. The vertically stacked panels report Hit@1, Hit@5, and Hit@10 within one unified table. Macro is the unweighted mean across datasets. All outputs share one disease normalizer; bold marks panel-best values.

(a) Hit@1: first-ranked diagnosis

Model	MyGene2	RAMEDIS	MME	HMS	LIRICAL	RDS	RDC	Pheno.	Macro
Closed-source LLMs									
Claude Opus 4.7	9.6	21.1	12.5	22.2	23.4	12.4	15.2	25.6	17.8
GLM-5.2	11.6	18.5	28.3	21.3	26.1	11.6	15.5	22.8	19.5
GPT-5.5	20.5	21.5	**29.5**	23.8	**27.2**	10.7	12.1	30.8	22.0
Open-weight baselines									
MedGemma1.5-4B	4.1	13.3	0.0	9.1	4.1	6.9	9.3	12.7	7.4
Qwen3-8B	5.5	12.8	0.0	18.2	9.7	4.8	6.7	5.0	7.8
Qwen3.5-9B	8.2	17.5	2.5	8.0	9.5	9.4	13.3	7.8	9.5
Qwen3.6-27B	8.2	20.4	7.5	20.4	12.7	12.5	15.6	12.0	13.7
Qwen3.6-35B-A3B	6.2	**22.1**	25.0	23.3	24.6	12.6	16.0	13.6	17.9
Occamy-1.0	6.2	3.2	0.0	10.2	5.1	1.7	2.1	7.2	4.5
Qwen3.6-35B-A3B (non-thinking)	8.2	15.5	10.0	18.2	13.0	14.8	19.5	10.2	13.7
Qwen3.8-27B	6.2	20.8	7.5	**27.3**	13.0	12.8	16.3	11.2	14.4
RareDx systems									
RareDx-9B, Direct (Ours)	**21.9**	10.4	5.0	12.5	21.4	10.4	17.6	**36.0**	16.9
RareDx-9B, Harness (Ours)	**21.9**	17.5	10.0	25.0	21.4	11.5	17.6	**36.0**	20.1
RareDx-27B, Direct (Ours)	15.8	9.9	5.0	12.5	21.4	16.0	**42.0**	**36.0**	19.8
RareDx-27B, Harness (Ours)	15.8	18.8	7.5	**27.3**	21.4	**19.4**	**42.0**	**36.0**	**23.5**

(b) Hit@5: diagnosis recovered within the top five

Model	MyGene2	RAMEDIS	MME	HMS	LIRICAL	RDS	RDC	Pheno.	Macro
Closed-source LLMs									
Claude Opus 4.7	30.8	37.7	29.2	36.5	**35.8**	22.3	23.6	38.0	31.7
GLM-5.2	22.6	33.8	32.6	37.7	35.0	21.7	24.9	30.6	29.9
GPT-5.5	**33.6**	33.4	32.8	36.2	35.0	18.7	20.3	**42.7**	31.6
Open-weight baselines									
MedGemma1.5-4B	17.1	21.0	0.0	20.5	8.6	15.5	18.2	30.9	16.5
Qwen3-8B	6.2	13.5	0.0	18.2	9.7	4.8	6.7	17.2	9.5
Qwen3.5-9B	20.5	37.3	7.5	23.9	16.5	19.6	23.4	16.8	20.7
Qwen3.6-27B	21.2	31.2	10.0	**46.6**	22.2	22.4	25.0	17.2	24.5
Qwen3.6-35B-A3B	16.4	36.0	**33.3**	36.7	35.1	20.8	22.9	20.2	27.7
Occamy-1.0	18.5	8.5	2.5	14.8	10.8	7.6	10.5	17.4	11.3
Qwen3.6-35B-A3B (non-thinking)	12.3	37.2	10.0	38.6	20.8	25.1	28.8	22.0	24.4
Qwen3.8-27B	23.3	31.9	10.0	39.8	20.0	24.2	26.9	21.4	24.7
RareDx systems									
RareDx-9B, Direct (Ours)	28.8	15.7	10.0	15.9	30.8	24.8	37.5	40.8	25.5
RareDx-9B, Harness (Ours)	28.8	37.7	10.0	39.8	30.3	27.1	37.8	40.8	31.5
RareDx-27B, Direct (Ours)	26.7	15.6	10.0	15.9	30.3	30.5	**63.4**	40.8	29.1
RareDx-27B, Harness (Ours)	26.7	**40.7**	10.0	42.0	30.3	**38.6**	**63.4**	40.8	**36.6**

(c) Hit@10: diagnosis recovered within the full differential

Model	MyGene2	RAMEDIS	MME	HMS	LIRICAL	RDS	RDC	Pheno.	Macro
Closed-source LLMs									
Claude Opus 4.7	34.2	41.2	**58.3**	41.3	**40.9**	26.5	28.5	43.2	39.3
GLM-5.2	30.8	**47.6**	39.1	41.0	39.0	26.7	29.3	34.4	36.0
GPT-5.5	39.0	45.1	37.7	40.0	37.8	23.6	24.7	**46.0**	36.7
Open-weight baselines									
MedGemma1.5-4B	23.3	22.8	2.5	23.9	11.1	19.9	22.2	38.2	20.5
Qwen3-8B	6.2	13.5	0.0	18.2	9.7	4.8	6.7	25.6	10.6
Qwen3.5-9B	25.3	39.6	10.0	33.0	19.7	23.3	26.7	20.6	24.8
Qwen3.6-27B	28.1	41.4	12.5	**52.3**	26.2	26.2	28.1	28.2	30.4
Qwen3.6-35B-A3B	25.3	41.9	41.7	40.0	40.2	24.7	26.4	23.6	33.0
Occamy-1.0	26.0	10.7	2.5	15.9	14.6	11.2	16.7	19.6	14.7
Qwen3.6-35B-A3B (non-thinking)	17.8	41.3	10.0	47.7	24.9	29.2	32.7	27.6	28.9
Qwen3.8-27B	25.3	36.4	10.0	47.7	23.0	28.8	30.6	26.4	28.5
RareDx systems									
RareDx-9B, Direct (Ours)	**41.1**	19.4	10.0	17.0	33.0	37.4	48.5	44.4	31.4
RareDx-9B, Harness (Ours)	**41.1**	42.3	10.0	50.0	31.9	37.8	49.2	44.4	38.3
RareDx-27B, Direct (Ours)	27.4	18.9	10.0	17.0	31.9	30.5	**63.6**	44.4	30.5
RareDx-27B, Harness (Ours)	27.4	45.7	12.5	51.1	32.1	**49.3**	**63.6**	44.4	**40.8**

Notes. Pheno. denotes Phenopackets. Models are ordered consistently within each panel. RareDx model rows use Top-10 SFT followed by RareDx-KGPO. Direct uses single-pass inference; Harness applies the archived exploratory dataset-level route.

aware term ($w = 0.3$) scores coherent, reversed, and unrelated lists at 1.000/0.997/0.911, whereas hit-only reward ties them. Constructions, intervals, and sensitivity curves are in Appendix J.

3.4 HARNESS COMPARISON AND COMPONENT ABLATION

Figure 4 shows gains of 3.21/6.00/6.99 points at Hit@1/5/10 for 9B and 3.71/7.41/10.30 for 27B. Larger gains at deeper cutoffs indicate improved differential coverage, but the effect is dataset-dependent, supporting case-aware rather than uniform evidence acquisition.

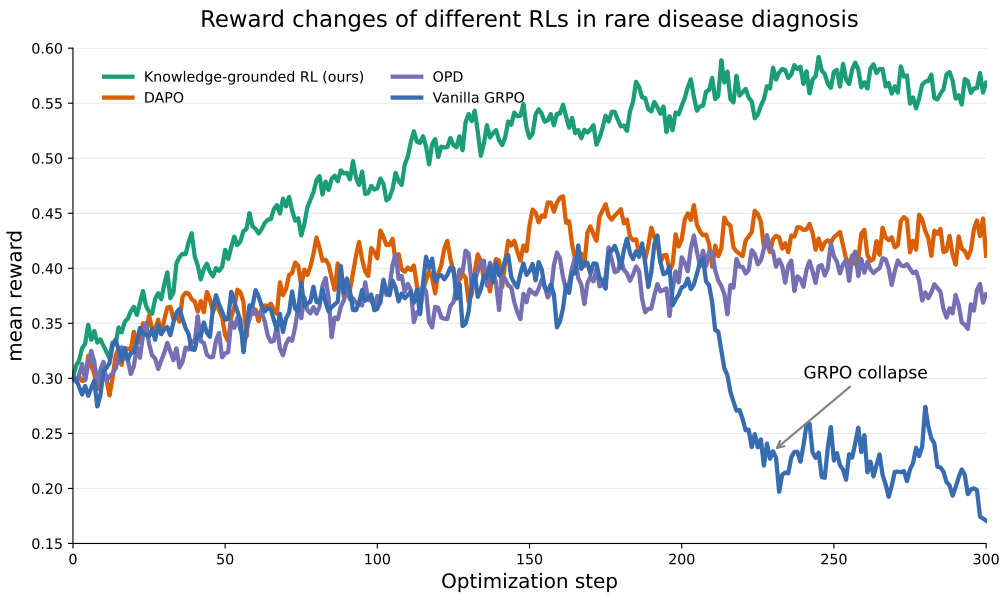

Figure 3: Reward dynamics for rare-disease post-training from a shared initialization. Vanilla GRPO improves initially but collapses after approximately 210 steps, consistent with reward hacking through indiscriminate disease-name generation. DAPO and OPD avoid catastrophic collapse but converge to lower rewards. RareDx-KGPO (Ours) supports sustained improvement and remains stable at the final checkpoint without late-stage collapse.

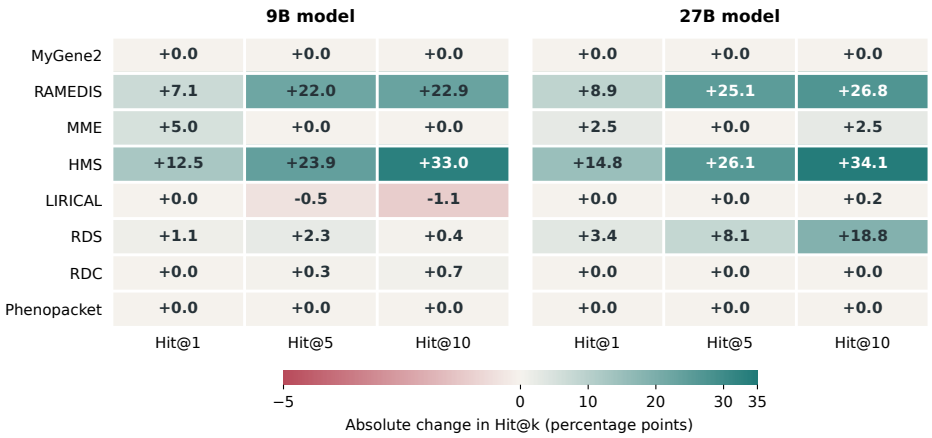

Figure 4: Per-dataset effect of applying the archived exploratory RareDx-Harness route to the same knowledge-grounded checkpoint. Cells show the absolute change in Hit@1, Hit@5, and Hit@10. Gains concentrate on RAMEDIS and HMS at both scales, with an additional large 27B gain on RDS. Near-zero cells correspond to retaining Direct when the audit provided no measured benefit; the small LIRICAL regression illustrates that routing is not error-free.

Static RAG underperforms Direct, showing that evidence volume is not evidence quality. Table 2 shows that ReAct widens coverage but slightly reduces Hit@1. On Phenopackets, HPO-Resnik raises gene recall@8 from 15.3% to 69.4% and Hit@10 from 27.5% to 36.3%, whereas dense gene retrieval adds less than one point. Structured retrieval can nevertheless introduce hard negatives elsewhere, explaining the need for routing. RRF should therefore be interpreted as variance reduction across stochastic lists, not as a remedy for systematically misleading evidence. The router is useful precisely because no acquisition strategy dominates across the heterogeneous records and evidence regimes represented by these benchmarks in our evaluation.

Table 2: Component ablations with fixed backbone and judge. Panel A uses Qwen3.6-Flash and macro averages over seven benchmarks. Panel B uses Qwen3.6-35B-A3B on Phenopackets.

Panel	Configuration	Hit@1	Hit@5	Hit@10
A	Direct	17.03	26.63	31.27
A	Adaptive ReAct	16.49	27.51	33.53
B	Direct	17.2	23.7	27.5
B	3-hop, dense gene retrieval	16.2	23.6	28.4
B	3-hop, HPO-Resnik	**22.2**	**31.7**	**36.3**

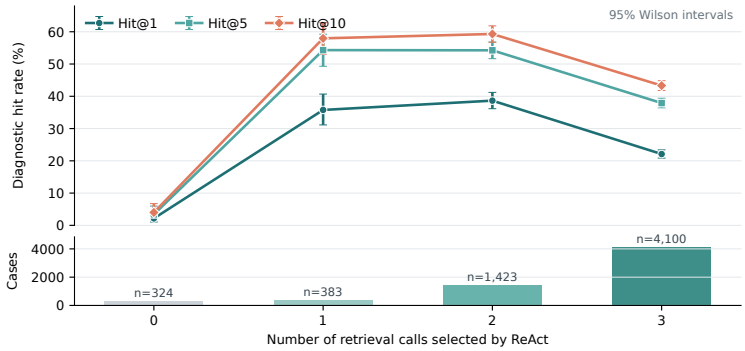

Figure 5: Observed accuracy by retrieval depth for 6,230 ReAct trajectories, with 95% Wilson intervals and stratum sizes. Depth is policy-selected, so the comparison describes behavior rather than estimating the causal effect of retrieval depth.

An adaptive-depth audit further finds a non-monotonic association between tool calls and Hit@10: one and two calls score 57.96 and 59.31, versus 43.32 at three calls. Because depth is policy-selected, this is descriptive rather than causal. Appendix K reports the complete strata, confidence intervals, and depth-conditioned diagnostic results.

3.5 Case study: from recognition to a usable differential

We audit the first 50 RAMEDIS records as a fixed slice using matched prompts, greedy decoding, token budget, parser, and normalizer. Direct 27B exceeds direct 9B at Hit@10, but the 9B Harness reverses the comparison and reaches 14/38/46 at Hit@1/5/10 (Table 3). This paired audit diagnoses system behavior; it does not replace the eight-benchmark result.

Table 3: Paired audit on the first 50 RAMEDIS records. Values are percentages under the disease matcher and parsing contract used throughout.

System	Hit@1	Hit@5	Hit@10
Qwen3.8-27B, direct	2	12	28
Ours 9B, direct	4	12	16
Ours 9B, full Harness	**14**	**38**	**46**

The 9B system recovers 16 diagnoses missed by direct 27B and loses seven, a net gain of nine Hit@10 cases; 13 recoveries enter the Top-5 rather than merely extending the tail. Under the same 1,024-token budget, direct 27B averages 462 words and never completes the required answer block, whereas the post-trained 9B model averages 91 words and completes all 50. Appendix I provides parsing details and a clinical example in which the Harness converts broad fatty-acid-oxidation recognition into the correct MCADD ranking. In that case, hypoglycemia, gastrointestinal symptoms, elevated transaminases, abnormal carnitine, and infantile death support a fatty-acid-oxidation disorder. The larger direct model recognizes the mechanism but does not produce a usable ranked answer before truncation. RareDx ranks MCADD first and organizes the remaining differential

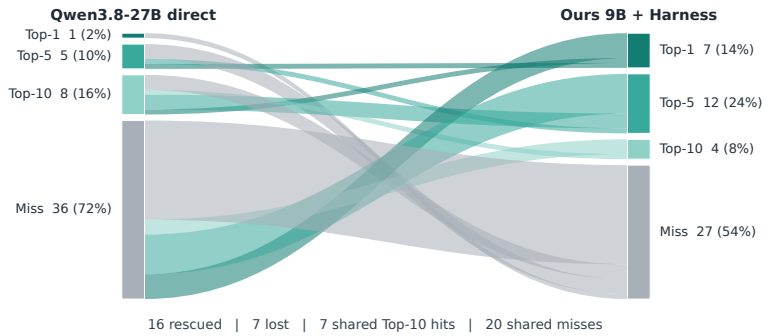

Figure 6: Paired rank-category transitions on the fixed 50-case RAMEDIS audit. Each ribbon follows one patient from direct Qwen3.8-27B to RareDx-9B with the full Harness. RareDx recovers 16 diagnoses missed by the larger model, 13 of which enter the Top-5.

around related oxidation and carnitine-transport disorders, illustrating how the output contract and controlled evidence turn broad recognition into an actionable ranking.

3.6 Efficiency analysis

On the same eight-A800 setup and 6,249 records, direct 9B and 27B generation takes 137.2 and 340.0 seconds, or 45.5 versus 18.4 cases/s, giving 9B a 2.48-fold throughput advantage. These timings cover direct batched generation and exclude loading, scoring, retrieval, and tool latency. Structured retrieval adds local computation without another model call; ReAct adds variable model-mediated turns. Appendix Table 16 reports the complete computational profile of every strategy.

4 Discussion

Rare-disease diagnosis benefits from structured knowledge, but retrieval is not monotonic: static context and high-recall gene candidates can introduce persuasive hard negatives, making routing part of the method. Conversely, ontology, phenotype, and curated relevance provide dense supervision while preserving entity validity and rank priority. Together these choices let RareDx-9B exceed GPT-5.5 at macro Hit@10 under the archived protocol and let RareDx-27B exceed it at every cutoff. The paired RAMEDIS audit confirms a system effect: direct 9B trails direct 27B, whereas the full 9B system reverses the comparison. More broadly, the results distinguish knowledge availability from knowledge use: the same store may help or distract depending on acquisition and case representation. RareDx makes this distinction measurable through a common output contract and learnable through entity-grounded rewards. Its 9B advantage therefore reflects controlled evidence acquisition, normalization, and ranking rather than a claim of greater parametric medical knowledge. Reliable interfaces to structured resources can rival backbone scale in long-tail tasks, especially at Hit@5 and Hit@10, where coherent coverage matters alongside first-choice confidence.

Limitations include dataset heterogeneity, the small MME split, and automatic rather than clinical adjudication. Archived routes reflect exploratory development; disjoint-selection results are the controlled estimate. Routed 9B results may invoke a 35B audit model, and direct throughput excludes loading, retrieval, and tool latency. Reward trajectories characterize optimization but do not replace matched downstream evaluation of every checkpoint. Finally, benchmark accuracy does not establish clinical safety; prospective studies must assess calibration, harmful omission, evidence faithfulness, and robustness across care settings. Future work should learn patient-level routing exclusively from development data, evaluate calibrated abstention, and measure whether retrieved evidence supports the final rank rather than merely correlating with it. Prospective studies should also report subgroup performance and clinician revision rates, since a plausible but misplaced rare diagnosis may impose costs that Hit@k alone cannot capture.

USE OF LARGE LANGUAGE MODELS

We used a large-language-model-based coding assistant during the development of this work. The assistant supported code review and refactoring, figure generation, and improvements to the clarity and presentation of the manuscript. The authors specified the research questions, designed the methods and experiments, interpreted the results, and reviewed all assistant-supported changes. All reported results were obtained from the described experiments, and all citations were checked against identifiable source publications. The authors take full responsibility for the manuscript and did not use language models to fabricate experimental evidence, citations, or supporting claims.

ETHICS STATEMENT

RareDx is developed for academic research on computational rare-disease diagnosis and is not intended to provide clinical advice or replace evaluation by qualified clinicians. Model outputs may be incomplete, incorrect, biased, or unsupported, particularly for underrepresented conditions and patient populations. Accordingly, the system should not be used for diagnosis, treatment selection, or other clinical decisions without independent expert review and appropriate validation. Users are responsible for complying with applicable requirements for patient privacy, data governance, informed consent, and responsible disclosure. Derivative models and materially modified system configurations should be released under distinct version identifiers so that their provenance and relationship to the evaluated artifacts remain traceable. We encourage explicit documentation of model, knowledge-base, and evaluation versions in any subsequent research use.

REPRODUCIBILITY STATEMENT

We archive the task definitions, verifier configurations, reference annotations, per-run predictions, evaluation summaries, model matrix, and ablation data underlying the reported tables and figures. The numerical results and plots are generated programmatically from these artifacts using fixed evaluation scripts. The accompanying repository contains the benchmark and evaluation code, plotting scripts, and experiment configurations needed to reproduce the reported analyses. Training and implementation details, including hardware, decoding settings, optimization parameters, and checkpoint-selection criteria, are provided in the appendix. Codes can be found in `https://anonymous.4open.science/r/RAREDX`.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=3zKtaqxLhW`.

Emily Alsentzer, Michelle M. Li, Shilpa N. Kobren, Ayush Noori, Undiagnosed Diseases Network, Isaac S. Kohane, and Marinka Zitnik. Few shot learning for phenotype-driven diagnosis of patients with rare genetic diseases. *npj Digital Medicine*, 8(1):380, 2025. doi: 10.1038/s41746-025-01749-1. URL `https://doi.org/10.1038/s41746-025-01749-1`.

Joanna S. Amberger, Carol A. Bocchini, Alan F. Scott, and Ada Hamosh. OMIM.org: leveraging knowledge across phenotype–gene relationships. *Nucleic Acids Research*, 47(D1):D1038–D1043, 2019. doi: 10.1093/nar/gky1151.

Anthropic. Introducing Claude Opus 4.7, April 2026. URL `https://www.anthropic.com/news/claude-opus-4-7`. Accessed: 2026-09-22.

Haichao Chen, Zhengyun Zhao, Songchi Zhou, Shikai Hu, Jinyuan Wang, Ye Jin, Xianghong Jin, Yih Chung Tham, Xiaofei Wang, Weizhi Ma, et al. Rarearena: a comprehensive benchmark dataset unveiling the potential of large language models in rare disease diagnosis. *The Lancet Digital Health*, 2026a.

Wenhui Chen, Shiwen Cheng, Hao Dong, Chenda Duan, Ruixiang Feng, Zhong Guan, Boqiang Guo, Xueyuan Han, Haojie Hao, Liangmeng Huang, Zhelong Huang, Xinke Kong, Hongyu Li, Jiazheng Li, Junbo Li, Qingchuan Li, Yukun Lian, Chang Liu, Tianyu Liu, Zicheng Liu, Shuyi

Ouyang, Yijun Pan, Kunyu Shi, Xiaojun Tang, Bingquan Wang, Kesu Wang, Yuchen Wang, Sibo Wei, Sicong Xie, Xiaoying Xing, Yi Xu, Zhijun Xu, Hongwei Xue, Qingcheng Zeng, Di Zhang, Guannan Zhang, Haochen Zhang, Tianlong Zhang, Tianyu Zhao, Tianyu Zhao, Yanjun Zheng, Jialong Zhu, and Zijian Zou. Occamy-1.0: Open pareto-frontier 35b intelligence for co-work, 2026b. URL https://arxiv.org/abs/2609.11977.

Xi Chen, Huahui Yi, Mingke You, WeiZhi Liu, Li Wang, Hairui Li, Xue Zhang, Yingman Guo, Lei Fan, Gang Chen, et al. Enhancing diagnostic capability with multi-agents conversational large language models. *NPJ digital medicine*, 8(1):159, 2025.

Xuanzhong Chen, Xiaohao Mao, Qihan Guo, Lun Wang, Shuyang Zhang, and Ting Chen. Rarebench: can llms serve as rare diseases specialists? In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*, pp. 4850–4861, 2024.

Xuanzhong Chen, Ye Jin, Xiaohao Mao, Lun Wang, Shuyang Zhang, and Ting Chen. RareAgents: Autonomous multi-disciplinary team for rare disease diagnosis and treatment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 101–109, 2026c. doi: 10.1609/aaai.v40i1.36969. URL https://doi.org/10.1609/aaai.v40i1.36969.

Daniel Danis, Michael J Bamshad, Yasemin Bridges, Andrés Caballero-Oteyza, Pilar Cacheiro, Leigh C Carmody, Leonardo Chimirri, Jessica X Chong, Ben Coleman, Raymond Dalgleish, et al. A corpus of ga4gh phenopackets: Case-level phenotyping for genomic diagnostics and discovery. *Human Genetics and Genomics Advances*, 6(1), 2025.

Jagjit Singh Dhatterwal, Mahaveer Singh Naruka, and Kuldeep Singh Kaswan. Multi-agent system based medical diagnosis using particle swarm optimization in healthcare. In *2023 International Conference on Artificial Intelligence and Smart Communication (AISC)*, pp. 889–893. IEEE, 2023.

Dong Dong, Roger Yat-Nork Chung, Rufina HW Chan, Shiwei Gong, and Richard Huan Xu. Why is misdiagnosis more likely among some people with rare diseases than others? insights from a population-based cross-sectional study in china. *Orphanet journal of rare diseases*, 15(1):307, 2020.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. The Faiss library. *arXiv preprint arXiv:2401.08281*, 2024. doi: 10.48550/arXiv.2401.08281.

Yuanqi Du, Botao Yu, Tianyu Liu, Tony Shen, Junwu Chen, Jan G Rittig, Kunyang Sun, Yikun Zhang, Zhangde Song, Bo Zhou, et al. Accelerating scientific discovery with autonomous goal-evolving agents. *arXiv preprint arXiv:2512.21782*, 2025.

Tapan Ghosh. Artificial intelligence in rare disease diagnostics: Shortening the path to early detection. 2025.

GLM-5-Team, Aohan Zeng, Xin Lv, Zhenyu Hou, et al. GLM-5: from vibe coding to agentic engineering, 2026. URL https://arxiv.org/abs/2602.15763.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3(1):1–23, 2021. doi: 10.1145/3458754.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081):633–638, 2025. doi: 10.1038/s41586-025-09422-z. URL https://doi.org/10.1038/s41586-025-09422-z.

Yaron Gurovich, Yair Hanani, Omri Bar, Guy Nadav, Nicole Fleischer, Dekel Gelbman, Lina Basel-Salmon, Peter M. Krawitz, Susanne B. Kamphausen, Martin Zenker, Lynne M. Bird, and Karen W. Gripp. Identifying facial phenotypes of genetic disorders using deep learning. *Nature Medicine*, 25(1):60–64, 2019. doi: 10.1038/s41591-018-0279-0. URL https://doi.org/10.1038/s41591-018-0279-0.

Tzung-Chien Hsieh, Aviram Bar-Haim, Shahida Moosa, Nadja Ehmke, Karen W. Gripp, Jean Tori Pantel, Magdalena Danyel, Martin Atta Mensah, Denise Horn, Stanislav Rosnev, Nicole Fleischer, Guilherme Bonini, Alexander Hustinx, Alexander Schmid, Alexej Knaus, Behnam Javanmardi, Hannah Klinkhammer, Hellen Lesmann, Sugirthan Sivalingam, Tom Kamphans, Wolfgang Meiswinkel, Frédéric Ebstein, Elke Krüger, Sébastien Küry, Stéphane Bézieau, Axel Schmidt, Sophia Peters, Hartmut Engels, Elisabeth Mangold, Martina Kreiß, Kirsten Cremer, Claudia Perne, Regina C. Betz, Tim Bender, Kathrin Grundmann-Hauser, Tobias B. Haack, Matias Wagner, Theresa Brunet, Heidi Beate Bentzen, Luisa Averdunk, Kimberly Christine Coetzer, Gholson J. Lyon, Malte Spielmann, Christian P. Schaaf, Stefan Mundlos, Markus M. Nöthen, and Peter M. Krawitz. GestaltMatcher facilitates rare disease matching using facial phenotype descriptors. *Nature Genetics*, 54(3):349–357, 2022. doi: 10.1038/s41588-021-01010-x. URL https://doi.org/10.1038/s41588-021-01010-x.

Karthik A. Jagadeesh, Johannes Birgmeier, Harendra Guturu, Cole A. Deisseroth, Aaron M. Wenger, Jonathan A. Bernstein, and Gill Bejerano. Phrank measures phenotype sets similarity to greatly improve mendelian diagnostic disease prioritization. *Genetics in Medicine*, 21(2): 464–470, 2019. doi: 10.1038/s41436-018-0072-y. URL https://doi.org/10.1038/s41436-018-0072-y.

Deyang Jiang, Haoran Wu, Ziyi Wang, Yiming Rong, Yunlong Zhao, Ye Jin, and Bo Xu. Raredxr1: Autonomous medical reasoning for rare disease diagnosis beyond human annotation. *arXiv preprint arXiv:2607.00147*, 2026.

Qiao Jin, Won Kim, Qingyu Chen, Donald C. Comeau, Lana Yeganova, W. John Wilbur, and Zhiyong Lu. MedCPT: Contrastive pre-trained transformers with large-scale PubMed search logs for zero-shot biomedical information retrieval. *Bioinformatics*, 39(11):btad651, 2023. doi: 10.1093/bioinformatics/btad651.

Anneliene H Jonker, Maria Cavaller-Bellaubi, Yukiko Nishimura, and David A Pearce. Access in the rare diseases landscape. *The Lancet Global Health*, 12(10):e1587, 2024.

Sebastian Köhler, Marcel H. Schulz, Peter Krawitz, Sebastian Bauer, Sandra Dölken, Claus E. Ott, Christine Mundlos, Denise Horn, Stefan Mundlos, and Peter N. Robinson. Clinical diagnostics in human genetics with semantic similarity searches in ontologies. *The American Journal of Human Genetics*, 85(4):457–464, 2009. doi: 10.1016/j.ajhg.2009.09.003. URL https://doi.org/10.1016/j.ajhg.2009.09.003.

Sebastian Köhler, Michael Gargano, Nicolas Matentzoglu, Leigh C. Carmody, David Lewis-Smith, Nicole A. Vasilevsky, Daniel Danis, Ganna Balagura, Gareth Baynam, Amy M. Brower, Tiffany J. Callahan, Christopher G. Chute, Johanna L. Est, Peter D. Galer, Shiva Ganesan, Matthias Griese, Matthias Haimel, Julia Pazmandi, Marc Hanauer, Nomi L. Harris, Michael J. Hartnett, Maximilian Hastreiter, Fabian Hauck, Yongqun He, Tim Jeske, Hugh Kearney, Gerhard Kindle, Christoph Klein, Katrin Knoflach, Roland Krause, David Lagorce, Julie A. McMurry, Jillian A. Miller, Monica C. Munoz-Torres, Rebecca L. Peters, Christina K. Rapp, Ana M. Rath, Shahmir A. Rind, Avi Z. Rosenberg, Michael M. Segal, Markus G. Seidel, Damian Smedley, Tomer Talmy, Yarlalu Thomas, Samuel A. Wiafe, Julie Xian, Zafer Yüksel, Ingo Helbig, Christopher J. Mungall, Melissa A. Haendel, and Peter N. Robinson. The Human Phenotype Ontology in 2021. *Nucleic Acids Research*, 49(D1):D1207–D1217, 2021. doi: 10.1093/nar/gkaa1043.

Junghwan Lee, Cong Liu, Junyoung Kim, Zhehuan Chen, Yingcheng Sun, James R Rogers, Wendy K Chung, and Chunhua Weng. Deep learning for rare disease: A scoping review. *Journal of biomedical informatics*, 135:104227, 2022.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.

Qigang Li, Keyan Zhao, Carlos D. Bustamante, Xin Ma, and Wing H. Wong. Xrare: a machine learning method jointly modeling phenotypes and genetic evidence for rare disease diagnosis.

Genetics in Medicine, 21(9):2126–2134, 2019. doi: 10.1038/s41436-019-0439-8. URL https://doi.org/10.1038/s41436-019-0439-8.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/aca97732e30bcf1303bc22ac3924fd16-Paper-Conference.pdf.

Chunyu Liu, Yixiao Jin, Zhouyu Guan, Tingyao Li, Yiming Qin, Bo Qian, Zehua Jiang, Yilan Wu, Xiangning Wang, Ying Feng Zheng, et al. Visual–language foundation models in medicine. *The Visual Computer*, 41(4):2953–2972, 2025a.

Tianyu Liu, Yijia Xiao, Xiao Luo, Hua Xu, Wenjin Zheng, and Hongyu Zhao. Geneverse: A collection of open-source multimodal large language models for genomic and proteomic research. In *Findings of the association for computational linguistics: EMNLP 2024*, pp. 4819–4836, 2024.

Tianyu Liu, Weihao Xuan, Hao Wu, Peter Humphrey, Marcello DiStasio, Heli Qi, Rui Yang, Simeng Han, Tinglin Huang, Fang Wu, et al. Teampath: Building multimodal pathology experts with reasoning ai copilots. *arXiv preprint arXiv:2511.17652*, 2025b.

Tianyu Liu, Tinglin Huang, Tong Ding, Hao Wu, Peter Humphrey, Sudhir Perincheri, Kurt Schalper, Rex Ying, Hua Xu, James Zou, et al. Leveraging multi-modal foundation models for analysing spatial multi-omic and histopathology data. *Nature Biomedical Engineering*, pp. 1–18, 2026a.

Tianyu Liu, Wangjie Zheng, Rui Yang, Benny Kai Guo Loo, Hui Zhang, Jeffries Lauran, Jianlei Gu, Botao Yu, Weihao Xuan, Kexin Huang, et al. A versatile ai agent for rare disease diagnosis and risk gene prioritization. *arXiv preprint arXiv:2605.06226*, 2026b.

Dongxue Mao, Chaozhong Liu, Linhua Wang, Rami Ai-Ouran, Cole Deisseroth, Sasidhar Pasupuleti, Seon Young Kim, Lucian Li, Jill A Rosenfeld, Linyan Meng, et al. Ai-marrvel—a knowledge-driven ai system for diagnosing mendelian disorders. *Nejm Ai*, 1(5):AIoa2300009, 2024.

Daniel McDuff, Mike Schaekermann, Tao Tu, Anil Palepu, Amy Wang, Jake Garrison, Karan Singhal, Yash Sharma, Shekoofeh Azizi, Kavita Kulkarni, Le Hou, Yong Cheng, Yun Liu, S. Sara Mahdavi, Sushant Prakash, Anupam Pathak, Christopher Semturs, Shwetak Patel, Dale R. Webster, Ewa Dominowska, Juraj Gottweis, Joelle Barral, Katherine Chou, Greg S. Corrado, Yossi Matias, Jake Sunshine, Alan Karthikesalingam, and Vivek Natarajan. Towards accurate differential diagnosis with large language models. *Nature*, 642(8067):451–457, 2025. doi: 10.1038/s41586-025-08869-4. URL https://doi.org/10.1038/s41586-025-08869-4.

Stéphanie Nguengang Wakap, Deborah M Lambert, Annie Olry, Charlotte Rodwell, Charlotte Gueydan, Valérie Lanneau, Daniel Murphy, Yann Le Cam, and Ana Rath. Estimating cumulative point prevalence of rare diseases: analysis of the orphanet database. *European journal of human genetics*, 28(2):165–173, 2020.

OpenAI. GPT-5.5 model, 2026. URL https://developers.openai.com/api/docs/models/gpt-5.5. API model documentation. Accessed: 2026-09-22.

Friedemann Paul. Hope for a rare disease: eculizumab in neuromyelitis optica. *The Lancet Neurology*, 12(6):529–531, 2013.

Ana Rath, Annie Olry, Ferdinand Dhombres, Maja Miličić Brandt, Bruno Urbero, and Ségolène Aymé. 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, 2012. doi: 10.1002/humu.22078.

Philip Resnik. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, pp. 448–453, 1995.

Peter N Robinson, Vida Ravanmehr, Julius OB Jacobsen, Daniel Danis, Xingmin Aaron Zhang, Leigh C Carmody, Michael A Gargano, Courtney L Thaxton, Guy Karlebach, Justin Reese, et al. Interpretable clinical genomics with a likelihood ratio paradigm. *The American Journal of Human Genetics*, 107(3):403–417, 2020.

Abeed Sarker, Rui Zhang, Yanshan Wang, Yunyu Xiao, Sudeshna Das, Dalton Schutte, David Oniani, Qianqian Xie, and Hua Xu. Natural language processing for digital health in the era of large language models. *Yearbook of Medical Informatics*, 33(01):229–240, 2024.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Damian Smedley, Julius OB Jacobsen, Marten Jäger, Sebastian Köhler, Manuel Holtgrewe, Max Schubach, Enrico Siragusa, Tomasz Zemojtel, Orion J Buske, Nicole L Washington, et al. Next-generation diagnostics and disease-gene discovery with the exomiser. *Nature protocols*, 10(12): 2004–2015, 2015.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. Large language models in medicine. *Nature medicine*, 29(8):1930–1940, 2023.

Khanh-Tung Tran, Dung Dao, Minh-Duong Nguyen, Quoc-Viet Pham, Barry O'Sullivan, and Hoang D Nguyen. Multi-agent collaboration mechanisms: A survey of llms. *arXiv preprint arXiv:2501.06322*, 2025.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems*, volume 36, pp. 74952–74965, 2023. doi: 10.52202/075280-3275. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/ed3fea9033a80fea1376299fa7863f4a-Paper-Conference.pdf.

Rodolfo Valdez, Lijing Ouyang, and Julie Bolen. Public health and rare diseases: oxymoron no more. *Preventing chronic disease*, 13:E05, 2016.

Nicole A. Vasilevsky, Sabrina Toro, Nicolas Matentzoglu, Joseph E. Flack, Kathleen R. Mullen, Harshad Hegde, Sarah Gehrke, Patricia L. Whetzel, Yousif Shwetar, Nomi L. Harris, Mee S. Ngu, Gioconda L. Alyea, Megan S. Kane, Paola Roncaglia, Eric Sid, Courtney L. Thaxton, Valerie Wood, Roshini S. Abraham, Maria Isabel Achatz, Pamela Ajuyah, Joanna S. Amberger, et al. Mondo: integrating disease terminology across communities. *Genetics*, 232(4):iyaf215, 2026. doi: 10.1093/genetics/iyaf215.

Hengchang Wang, Li Liu, Huaxiang Zhang, Lei Zhu, Xiaojun Chang, and Hao Du. Visualrag: Knowledge-guided retrieval augmentation for image-text matching. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2203.11171.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837, 2022. doi: 10.52202/068431-1800. URL https://proceedings.neurips.cc/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. Benchmarking retrieval-augmented generation for medicine. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 6233–6251, 2024. doi: 10.18653/v1/2024.findings-acl.372. URL https://aclanthology.org/2024.findings-acl.372/.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2210.03629.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025.

Tomasz Zemojtel, Sebastian Köhler, Luisa Mackenroth, Marten Jäger, Jochen Hecht, Peter Krawitz, Luitgard Graul-Neumann, Sandra Doelken, Nadja Ehmke, Malte Spielmann, Nancy Christine Øien, Michal R. Schweiger, Ulrike Krüger, Götz Frommer, Björn Fischer, Uwe Kornak, Ricarda Flöttmann, Amin Ardeshirdavani, Yves Moreau, Suzanna E. Lewis, Melissa Haendel, Damian Smedley, Denise Horn, Stefan Mundlos, and Peter N. Robinson. Effective diagnosis of genetic disease by computational phenotype analysis of the disease-associated genome. *Science Translational Medicine*, 6(252):252ra123, 2014. doi: 10.1126/scitranslmed.3009262. URL https://doi.org/10.1126/scitranslmed.3009262.

Mengge Zhao, James M Havrilla, Li Fang, Ying Chen, Jacqueline Peng, Cong Liu, Chao Wu, Mahdi Sarmady, Pablo Botas, Julián Isla, Gholson J Lyon, Chunhua Weng, and Kai Wang. Phen2Gene: rapid phenotype-driven gene prioritization for rare diseases. *NAR Genomics and Bioinformatics*, 2(2):lqaa032, 2020. doi: 10.1093/nargab/lqaa032. URL https://doi.org/10.1093/nargab/lqaa032.

Weike Zhao, Chaoyi Wu, Yanjie Fan, Pengcheng Qiu, Xiaoman Zhang, Yuze Sun, Xiao Zhou, Shuju Zhang, Yu Peng, Yanfeng Wang, et al. An agentic system for rare disease diagnosis with traceable reasoning. *Nature*, 651(8106):775–784, 2026a.

Weike Zhao, Chaoyi Wu, Yanjie Fan, Xiaoman Zhang, Pengcheng Qiu, Yuze Sun, Xiao Zhou, Yanfeng Wang, Ya Zhang, Yongguo Yu, et al. An agentic system for rare disease diagnosis with traceable reasoning. *Nature*, 2026b.

A Related Work

A.1 Rare disease diagnosis

Phenotype-driven diagnostic methods provide a foundation for matching incomplete clinical presentations to candidate disorders. Phenomizer uses ontology-based semantic similarity and statistical significance to rank diseases, accounting for noisy and imprecisely described phenotypes (Köhler et al., 2009). PhenIX integrates semantic phenotype matching with variant pathogenicity analysis in a sequencing-based diagnostic workflow (Zemojtel et al., 2014). Phrank uses an information-theoretic, Bayesian-network formulation to compare phenotype sets and prioritize diseases or genes independently of the underlying knowledge base and variant-filtering scheme (Jagadeesh et al., 2019). Phen2Gene takes a complementary approach: it aggregates weighted phenotype-to-gene associations from a precomputed knowledge base to produce candidate-gene rankings from HPO terms (Zhao et al., 2020) for early diagnosis.

These approaches establish that the specificity and relationships of phenotypic findings can be more informative than exact term overlap. They also expose distinct diagnostic endpoints: ranking diseases, prioritizing genes, and identifying causative variants require different inputs and validation criteria. RareDx builds on this structured-knowledge tradition but targets generative disease ranking across heterogeneous records. Its canonicalization and phenotype-grounded reward transfer curated relationships into supervision for a bounded differential. This use differs from applying a phenotype similarity score directly as the inference-time ranking function. Moreover, partial credit for a related disease is a training signal rather than a calibrated probability that the patient has that disease.

A.2 AI for rare disease diagnosis

Learning-based systems extend rare-disease analysis through specialized representations and the integration of multiple evidence sources. Xrare combines phenotype similarity with genetic evidence and guideline-derived features in a gradient-boosted model for causative variant prioritization (Li et al., 2019). SHEPHERD uses geometric deep learning over a biomedical knowledge graph to learn patient representations with scarce diagnostic labels, supporting causal-gene prioritization, similar-patient retrieval, and disease characterization (Alsentzer et al., 2025). Both illustrate how structured biomedical knowledge can inform learning when individual disorders have limited labeled examples. Their prediction objects and supervision also differ from optimizing and tuning LLMs to generate a list of diseases.

Visual phenotyping supplies another complementary source of evidence. DeepGestalt learns facial representations for genetic-syndrome recognition (Gurovich et al., 2019), while GestaltMatcher places facial phenotype descriptors in a shared space to match patients, including individuals with syndromes unseen during training (Hsieh et al., 2022). These methods are relevant to the broader diagnostic workflow, but their image-based setting does not provide a direct performance comparator for phenotype-text diagnosis.

LLM-based agents instead organize knowledge acquisition and specialist interaction. RareAgents, Hygieia, and DeepRare combine multidisciplinary coordination, memory, and medical tools for rare-disease diagnosis and treatment support (Chen et al., 2026c; Liu et al., 2026b; Zhao et al., 2026a). Such systems motivate evaluating the contribution of orchestration alongside that of the underlying model. RareDx addresses this question through a shared evidence layer, output contract, and disease normalizer for its inference-strategy comparisons, together with graph-grounded post-training. Its final operational router selects between Direct and Diagnostic Audit at the dataset level, which is distinct from the broader retrieval and tool configurations studied in the harness. These comparisons position the method without implying superiority over systems evaluated with different inputs, backbones, or tool access.

A.3 LLMs and reasoning

Chain-of-thought prompting elicits intermediate reasoning through worked examples (Wei et al., 2022), and self-consistency samples multiple reasoning paths before aggregating their answers (Wang et al., 2023). Retrieval-augmented generation combines parametric knowledge with an external retrieval index (Lewis et al., 2020), whereas ReAct interleaves reasoning with actions

that acquire additional information (Yao et al., 2023). These approaches motivate separating direct generation, static retrieval, adaptive tool use, and aggregation in a diagnostic harness. RareDx's reciprocal-rank fusion aggregates disease lists rather than applying majority voting to a single answer, so its output object differs from standard self-consistency.

Medical evaluations further show why evidence use must be studied in context. MedRAG and the MIRAGE benchmark systematically compare medical corpora, retrievers, and backbone models, and identify sensitivity to retrieval configuration and context placement (Xiong et al., 2024). AMIE evaluates LLM-generated differential diagnoses both alone and as assistance to clinicians on challenging published case reports (McDuff et al., 2025). These studies support examining retrieval design and differential quality, while medical question answering and general diagnostic case reports remain different evaluation settings from RareDx's rare-disease benchmarks. Explanation quality also requires separate scrutiny: Turpin et al. (2023) demonstrate that chain-of-thought explanations can omit influential biases and rationalize incorrect answers. Consequently, diagnostic ranking accuracy and the presence of a fluent reasoning trace should not be treated as evidence of faithful clinical reasoning.

Post-training introduces a further distinction between improving outcomes and supervising the reasoning process. DeepSeek-R1 demonstrates the development of reasoning behavior through large-scale reinforcement learning with verifiable feedback (Guo et al., 2025). Lightman et al. (2024) compare final-answer supervision with feedback on individual reasoning steps, finding advantages for process supervision in mathematical problem solving. RareDxR1, which is closed source, also leverages the contributions in DeepSeek-R1 and builds a reasoning model for rare disease diagnosis Jiang et al. (2026). Transferring these ideas to rare-disease diagnosis requires specifying what can be verified: canonical disease identity, ontology relationships, and phenotype consistency offer structured signals, but do not validate every intermediate clinical inference. RareDx-KGPO uses these signals to shape rewards for ranked diagnostic outputs, with vocabulary and list-budget safeguards against invalid candidates and indiscriminate enumeration.

B BENCHMARK COMPOSITION

The macro score gives each of the eight reported benchmark sources equal weight despite their different sample counts. RareBench is evaluated on all questions, with RAMEDIS, MME, HMS, and LIRICAL reported separately to expose source-level variation. Table 4 records the composition used throughout the main results.

Table 4: Number of test records contributing to the macro average. RareBench coverage includes all questions, reported separately for RAMEDIS, MME, HMS, and LIRICAL.

Dataset	Cases	Dataset	Cases
MyGene2	146	RAMEDIS	624
MME	40	HMS	88
LIRICAL	370	RDS	1,803
RDC	678	Phenopackets	500
		Total	4,249

C HARNESS ROUTER SPECIFICATION AND LEAKAGE-CONTROLLED AUDIT

The phrase "selected Harness strategy" in Table 1 refers to a *dataset-level* router, not a per-case learned gate. Its two executable candidates are (i) DIRECT, the frozen RareDx Top-10 response, and (ii) AUDIT, a deterministic diagnostic audit by a locally served Qwen3.6-35B-A3B model. The auditor receives the original patient evidence and the Direct ranked list, is instructed to retain supported entries and repair unsupported or missing diagnoses, and returns ten canonical disease names. It uses greedy decoding (temperature 0), a 384-token output cap, disabled thinking, and no retrieval or external tools. If parsing yields fewer than two audit candidates, the Direct list is retained; otherwise the audit list is deduplicated and any unfilled tail is copied from Direct. The four inference strate-

gies compared in the main Harness study are research configurations; the final router only chooses between the two archived candidates above.

To make route selection independent of the evaluation portion, we add a strict audit over frozen outputs. For every model-dataset pair, a SHA-256 hash of `seed|dataset|case-id` assigns approximately 20% of cases to development and the remainder to test (seed 20260919). Let $m_s^D = (H@10, H@5, H@1)$ be strategy s's development tuple. The router is

$$s_D^* = \arg\operatorname*{lex\,max}_{s \in \mathcal{S}_D} m_s^D, \qquad \mathcal{S}_D \subseteq \{\text{Direct}, \text{Audit}\}, \tag{8}$$

where the lexicographic comparison prioritizes Hit@10, then Hit@5 and Hit@1; an exact tie selects DIRECT, the cheaper strategy. The decision is made once per model and benchmark source, never per patient. Audit is included in $\mathcal{S}_D$ only when a complete frozen audit output exists for that source. No test score enters Equation 8.

Table 5: Validation-only route audit. Each metric cell is Hit@1/5/10 (%). D and A denote Direct and Diagnostic Audit. A dash means that a complete archived Audit candidate was not available, so the candidate set contained Direct only.

Model	Dataset	n_{dev}	n_{test}	D, dev	A, dev	Pick	D, test	A, test
9B	MyGene2	33	113	24.2/27.3/39.4	12.1/18.2/27.3	D	21.2/29.2/41.6	14.2/29.2/32.7
9B	RAMEDIS	113	511	11.5/18.6/21.2	16.8/40.7/47.8	A	9.6/14.9/18.4	17.6/37.0/41.1
9B	MME	9	31	0.0/11.1/11.1	0.0/0.0/0.0	D	6.5/9.7/9.7	12.9/12.9/12.9
9B	HMS	22	66	22.7/27.3/27.3	27.3/45.5/59.1	A	9.1/12.1/13.6	24.2/37.9/47.0
9B	LIRICAL	62	308	17.7/27.4/29.0	16.1/27.4/30.7	A	22.1/30.8/32.5	19.2/27.0/30.8
9B	RDS	342	1,461	11.7/27.2/38.3	–	D	10.1/24.2/37.2	–
9B	RDC	127	551	18.1/36.2/48.8	–	D	17.4/37.8/48.5	–
9B	Phenopackets	103	397	42.7/47.6/50.5	–	D	34.3/39.0/42.8	–
27B	MyGene2	33	113	9.1/24.2/24.2	6.1/21.2/21.2	D	17.7/27.4/28.3	14.2/26.6/29.2
27B	RAMEDIS	113	511	14.2/21.2/21.2	17.7/46.0/50.4	A	10.8/24.3/24.3	19.0/39.5/44.6
27B	MME	9	31	0.0/0.0/0.0	0.0/0.0/0.0	D	3.2/3.2/3.2	9.7/12.9/16.1
27B	HMS	22	66	36.4/36.4/36.4	31.8/54.6/68.2	A	22.7/27.3/27.3	25.8/37.9/45.5
27B	LIRICAL	62	308	16.1/19.4/19.4	16.1/29.0/29.0	A	14.9/21.1/21.1	16.2/27.0/29.9
27B	RDS	342	1,461	15.8/30.4/30.4	19.6/38.6/48.0	A	16.0/30.5/30.5	19.4/38.6/49.6
27B	RDC	127	551	40.2/58.3/59.1	26.8/47.2/57.5	D	42.5/64.6/64.6	29.8/48.3/64.4
27B	Phenopackets	103	397	43.7/44.7/44.7	–	D	30.0/39.3/39.3	–

Across the eight test partitions, the validation-only router obtains 18.81/30.21/37.34 for 9B and 21.71/34.69/38.13 for 27B. We also repeat the complete selection protocol for ten fixed hash seeds. Audit is selected in all ten splits for 9B RAMEDIS and HMS and for 27B RAMEDIS, LIRICAL, and RDS. Selection is less stable for the small MME set and for MyGene2, and it varies for 9B LIRICAL and 27B RDC. This sensitivity is why the strict audit reports every candidate rather than only the selected route.

The headline routing in Table 1 was produced earlier during exploratory development using benchmark-level results; it was not selected under the strict split protocol above. Consequently, Table 5, rather than the headline routed row, is the appropriate estimate when validation-only route selection is required. The frozen Direct rows remain unaffected by this distinction.

D EXACT REWARD IMPLEMENTATION

Let y_k be the disease string at rank $k \leq 10$. Normalization removes punctuation and parenthetical aliases and maps y_k to the nearest canonical vocabulary entry only when normalized Levenshtein distance is at most 0.20. An unmapped string receives no dense medical credit unless it is explicitly present in the curated graded-gold set. For a mapped candidate, the four reward components shown

in Figure 2 are

$$g_k = \max_{r \in \mathcal{C}} \gamma(r)\, \mathbb{1}[\mathrm{norm}(y_k) = r], \tag{9}$$

$$o_k = \max_{g \in G} 2^{-d_{\mathcal{O}}(y_k, g)}, \tag{10}$$

$$e_k = \max_{g \in G} f(\cos(z_{y_k}, z_g)), \tag{11}$$

$$p_k = \min\{0.40, \mathrm{WJaccard}_{IC}(P(y_k), P(y^*))\}, \tag{12}$$

$$c_k = \max\{g_k, o_k, e_k, p_k\}, \tag{13}$$

where $\mathcal{C}$ is the curated set containing the accepted diagnosis and clinically related neighbors, $\gamma(r) \in (0, 1]$ is their graded relevance, G is the accepted diagnosis set, $d_{\mathcal{O}}$ is shortest-path distance in the disease ontology, and $P(y^*)$ is the explicit reference HPO set when provided or otherwise the knowledge-graph profile of the reference disease. The symmetric phenotype overlap is weighted by HPO information content. The embedding transform is

$$f(s) = \begin{cases} 0, & s \le 0.75, \\ 0.30(s - 0.75)/0.05, & 0.75 < s < 0.80, \\ 0.30 + 0.05(s - 0.80)/0.10, & 0.80 \le s < 0.90, \\ 0.35, & s \ge 0.90. \end{cases} \tag{14}$$

Thus exact or synonymous ontology matches receive 1, while one- and two-edge relations receive 0.5 and 0.25. The rank-sensitive base score is $H = \max_{k \le 10} c_k / \log_2(k + 1)$.

The anti-hacking transform is applied after medical scoring. A candidate line is contaminated when it is degenerate and cannot be resolved to a real disease; degeneracy is triggered by more than 12 words, nonword-character mass above 0.5, or token repetition of at least 0.5. If m lines are contaminated, the cleanliness multiplier is 2^{-m}. Outputs with more than ten raw candidates receive a 0.9 budget multiplier. To protect a verified exact rank-1 answer from a residual false-positive contamination heuristic, the deployed configuration floors its cleanliness multiplier at 1. A valid anchor requires ontology distance at most one or embedding cosine at least 0.80. A response is globally invalid when at least half of its lines are degenerate or fewer than half are resolvable; a globally invalid response without an anchor receives -1, while a missing or empty `<answer>` block receives 0. For the reported checkpoint, define $q = 1$ for an exact ontology match at rank 1 and $q = 2^{-m}$ otherwise. In the general formulation illustrated in Figure 2,

$$R = \big[(1 - w)H + wN\big] T_{\mathrm{turn}} H_{\mathrm{hedge}} C_{\mathrm{clean}}. \tag{15}$$

For the reported single-turn run, $w = 0$, $T_{\mathrm{turn}} = 1$, $H_{\mathrm{hedge}} = 0.9^{\mathbb{1}[L > 10]}$, and $C_{\mathrm{clean}} = q$, giving

$$R = H\, q\, 0.9^{\mathbb{1}[L > 10]}, \tag{16}$$

subject to the malformed and globally invalid overrides above. The list-level nDCG extension is implemented in the analysis code but its weight is zero for the reported training run; it is used only in the fixed-output audit in Section J. This distinction prevents the audit-only term from being attributed to the trained checkpoint.

E Training provenance and overlap audit

Table 6 records the exact artifacts referenced by the 9B launch scripts. SFT converts each source answer into an ordered Top-10 target using seed 13, placing the accepted diagnosis first and filling alternatives from teacher predictions, KG phenotype neighbors, ontology neighbors, then lexical or random fallbacks when graph coverage is insufficient. The RL mixture totals 61,918 prompts. The RareArena article-backed portion contains recorded publication dates from 2003-08-18 through 2024-06-15 for 22,403 of 42,977 rows; the remaining artifacts do not store a reliable source-publication date, so no broader temporal cutoff is claimed. Train and validation are separate materialized files. The archived preparation code does not retain the original randomization manifest, which prevents reconstructing a stronger chronological split claim after the fact.

We performed a read-only overlap audit against all eight evaluation sources. Three signatures are checked independently: normalized patient/case identifier, SHA-1 of normalized case text, and exact

Table 6: Training and validation artifacts used by the reported 9B pipeline.

Stage	Component	Rows	Content
SFT	train	51,368	RareArena cases with constructed Top-10 targets
SFT	validation	932	held-out Phenopacket-format prompts
RL	RareArena	42,977	article-backed case reports and test results
RL	Phenopacket	8,392	phenotype/gene disease prompts
RL	gene-phenotype hard set	3,928	synthetic hard cases
RL	simulated HPO subset	2,714	synthetic incomplete-phenotype cases
RL	sparse/atypical set	3,907	synthetic sparse or atypical cases
RL	validation	932	Phenopacket-format prompts

or near-exact disease-plus-phenotype set (HPO Jaccard at least 0.90). There are no identifier or case-text matches in any SFT or RL training artifact. SFT has no phenotype-signature match. RL training has one exact disease-plus-phenotype signature shared with LIRICAL (1 of 61,918 prompts), without a shared identifier or case text. The RL validation file shares 290 disease-plus-phenotype profiles with the 500-case Phenopackets benchmark, again with no identifier or case-text match. These are profile-level matches and do not establish patient identity, but they can make Phenopackets-based checkpoint selection optimistic. As a conservative sensitivity analysis, removing Phenopackets from the macro average leaves RareDx-9B at 37.47 Hit@10 versus 35.41 for GPT-5.5, increasing rather than reversing the reported margin.

The retrieval collection is intentionally not entity-disjoint from evaluation: it contains 27,554 disease-level documents built from HPO, Orphanet, OMIM, and MONDO, plus an optional 2,299 gene documents. These records contain disease names, definitions, synonyms, and phenotype/gene associations but not benchmark patient reports. Entity and phenotype overlap is therefore the intended retrieval signal, whereas patient-text overlap is not. Source snapshot dates were not embedded in the archived indices, so we do not claim a temporal-disjointness guarantee for the retrieval corpus.

F POST-TRAINING PROMPT AND SERIALIZATION

We inspected every record referenced by the 27B RL launch entry. All 61,918 training records and 932 validation records contain exactly one `user` message, no stored `system` message, and the same instruction prefix. The five training components differ only in the case payload: article-derived reports may include free-text findings and diagnostic tests, whereas the Phenopacket-derived and synthetic components express observed phenotypes, explicitly absent phenotypes when available, and occasionally a candidate gene. The reference disease is stored in the reward metadata and is not inserted as a separate prompt field. At runtime, the Qwen chat template serializes this single message and adds the model-specific control tokens.

The following is the verbatim instruction template, with only the record-dependent clinical text replaced by a bracketed placeholder. Whitespace and the ten output slots match the materialized Parquet records.

```
You are an expert physician specialized in rare disease diagnosis.

Read the patient's phenotype description and reason step by step.

First put your full reasoning inside <think> </think>.

Then output your differential diagnosis as a ranked list of the 10 most
likely rare diseases, most likely first, inside <answer> </answer>.

Format strictly as:

<think> your reasoning </think>

<answer>

1.  disease name

2.  disease name
```

```
3.  disease name
4.  disease name
5.  disease name
6.  disease name
7.  disease name
8.  disease name
9.  disease name
10.  disease name
</answer>


Patient phenotypes:
```

[record-dependent case description, observed/absent phenotypes, genetic findings, and diagnostic tests when available]

The output contract serves two purposes. It fixes the evaluation budget to one ordered Top-10 differential and separates free-form reasoning from the machine-scored list. The reward parser matches the `<answer>` block case-insensitively, removes numbering and bullet syntax, and scores at most the first ten parsed candidates in rank order. A missing or empty answer block receives zero under the deployed reward; extra candidates trigger the output-budget multiplier described in Appendix D. The `<think>` block is parsed separately for the phenotype-grounding safeguard and is never treated as an additional disease list. Thus, the prompt, parser, reward budget, and Hit@1/5/10 evaluation expose the same ordered output object.

Audited 27B launch configuration. The launch initializes from the Top-10 SFT export at `global_step_401` and mixes the five RL components enumerated in Table 6. It uses 16 A800-80GB GPUs across two nodes, eight rollouts per prompt, training and generation batches of 256, and a PPO minibatch of 64. Prompts and responses are capped at 1,024 and 768 tokens. Training samples with temperature 1.0, top-$p = 0.95$, and top-$k = 50$; validation uses greedy decoding. The actor learning rate is 2×10^{-7}, the reference-policy KL coefficient is 0.15, and the lower/upper policy-ratio clips are 0.20/0.28. Zero-variance rollout groups are resampled. Validation runs every five steps and checkpoints are written every ten steps, enabling selection on the held-out 932-record validation file rather than on training reward. These values are read from the executable launch configuration rather than from comments or log-derived defaults.

G PROPRIETARY BASELINE PROTOCOL

All proprietary rows use the same task template: a clinical-geneticist instruction, the patient phenotypes or case report (and test results for RDC), and a requirement to return exactly ten specific formal disease names in descending order, one per line, without explanation. The system message is `You are a helpful assistant.` No retrieval, function call, web search, or other tool is enabled, and the intended protocol makes one generation per case with no self-consistency aggregation.

The notebooks retain aliases but not immutable provider snapshot identifiers, response headers, run timestamps, or provider-side default values. Therefore the exact serving revisions and numeric decoding defaults cannot be recovered and are not inferred here. This is a limitation of the original baseline collection. Future releases will persist request JSON, resolved model revision, response metadata, per-case attempt count, and raw output for every proprietary call.

H POST-TRAINING CONFIGURATIONS

Table 8 separates the optimization rule, supervision signal, and need for an online teacher. This distinction is important because our contribution is the medical graph reward rather than a new group-normalization update.

Table 7: Recorded proprietary-baseline invocation. "Provider default" means the argument is absent from the archived request rather than assigned an inferred value.

Reported row	Request alias	Interface	Decoding and tools
GPT-5.5	`gpt-5.5`	OpenAI-compatible proxy, `xiaoai.plus/v1`	temperature, top-p, seed, and output cap: provider default; tools: none
Claude Opus 4.7	`claude-opus-4-7`	OpenAI-compatible proxy, `xiaoai.plus/v1`	temperature, top-p, seed, and output cap: provider default; tools: none
GLM-5.2	`glm-5.2`	Alibaba MaaS compatible API for phenotype sets; DashScope Generation for RDS/RDC	decoding and output cap: provider default; tools: none

Table 8: Optimization and supervision differences in the RL comparison.

Method	Update	Signal	Teacher
GRPO	Group relative	Task reward	No
DAPO	Dynamic groups, asymmetric clip	Task reward	No
OPD	On-policy distillation	Token distributions	Yes
RareDx-KGPO (Ours)	Group relative	Medical graph reward	No

I CASE-STUDY DETAILS

The case study uses row 12 of the frozen RAMEDIS evaluation split. Its phenotype list is {death in infancy, hypoglycemia, vomiting, diarrhea, elevated hepatic transaminases, abnormal circulating carnitine concentration}. Both models receive the same system instruction and ranked-differential template, use greedy decoding, and have a maximum generation length of 1,024 tokens. The complete 9B Harness list is: medium-chain acyl-CoA dehydrogenase deficiency; carnitine palmitoyltransferase II deficiency; very-long-chain acyl-CoA dehydrogenase deficiency; multiple acyl-CoA dehydrogenase deficiency; carnitine-acylcarnitine translocase deficiency; primary systemic carnitine deficiency; mitochondrial trifunctional protein deficiency; short-chain acyl-CoA dehydrogenase deficiency; glutaric acidemia type II; and propionic acidemia.

For the 50-case audit, we take rows 0-49 before inspecting outputs. The direct 27B outputs are newly generated, while direct 9B and Harness outputs come from the frozen evaluation artifacts used in the main experiment. All three are rescored with the same parser and name matcher. The audit contains only RAMEDIS cases and is deliberately reported separately from the eight-dataset macro average.

Table 9: Observed outputs for RAMEDIS case 12.

Model	Observed output
Qwen3.8-27B	Restates phenotype categories and surveys metabolic mechanisms, but does not produce a disease-ranked answer before truncation. MCADD is not recovered.
Ours 9B + Harness	Ranks MCADD first, followed by carnitine palmitoyltransferase II deficiency, very-long-chain acyl-CoA dehydrogenase deficiency, multiple acyl-CoA dehydrogenase deficiency, and carnitine-acylcarnitine translocase deficiency.

J REWARD-AUDIT CONSTRUCTION

We select the first 120 validation diseases that resolve to the frozen disease graph and have at least one distinct disease sharing an HPO annotation. For the phenotype-neighbor probe, we choose the non-reference disease maximizing set Jaccard similarity between the two disease HPO profiles. The pseudo-disease probe prepends and appends diagnostic modifiers to the normalized reference name, creating a medically styled string that is absent from the canonical vocabulary. The over-budget

Table 10: Direct-generation behavior on the 50-case audit. Answer-block compliance requires both opening and closing tags. Output words are counted after whitespace tokenization and are reported only as a model-agnostic verbosity diagnostic.

Direct model	Complete block	Mean words	Median words
Qwen3.8-27B	0/50	462.2	451.0
Ours 9B	50/50	91.3	90.5

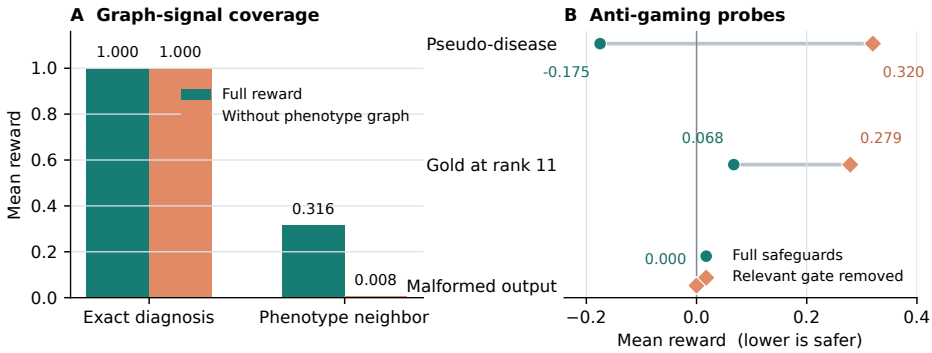

Figure 7: Controlled reward audit on 120 held-out diagnoses. Panel A removes the disease-phenotype graph signal while holding outputs fixed. Panel B removes the vocabulary gate for pseudo-diseases or the Top-10 budget for a delayed exact match.

probe places ten real distractor diseases before the reference. All prompts retain their original validation metadata. Biomedical name embeddings are disabled for both the full and ablated configurations in this audit, while the same graph, disease profiles, parser, and normalization thresholds are retained.

Table 11: Numerical reward audit. "Ablated" removes the phenotype-graph channel, vocabulary gate, or Top-10 budget for the corresponding probe. Positive is the percentage of constructed outputs receiving reward greater than zero.

Probe	Full	Ablated	Positive, full	Positive, ablated
Exact diagnosis	1.000	1.000	100.0	100.0
Phenotype neighbor	0.316	0.008	100.0	0.8
Pseudo-disease	-0.175	0.320	0.0	82.5
Gold at rank 11	0.068	0.279	100.0	100.0
Malformed output	0.000	0.000	0.0	0.0

J.1 List-quality discrimination audit

For each eligible validation disease, we rank other graph diseases by Jaccard overlap of their HPO profiles and retain the four strongest distinct neighbors. Their graded relevance is the observed overlap capped at 0.6; the reference receives 1.0. The coherent list places these neighbors in descending relevance after the rank-1 reference, the reverse list flips only their order, and the unrelated list replaces them with deterministic real-disease distractors. All configurations use the same parser, vocabulary gate, Top-10 budget, disease graph, and phenotype channel; embeddings are disabled to avoid cross-signal substitution. Confidence intervals use 2,000 case-level bootstrap resamples with a fixed seed. At $w = 0.3$, the unrelated-list mean is 0.911 (95% CI 0.907-0.914), while the coherent-list mean is 1.000. This experiment also exposed and corrected a normalization error: the DCG numerator must use the same curated relevance set as IDCG, rather than arbitrary dense similarity. The headline $w = 0$ checkpoints are unaffected by this correction.

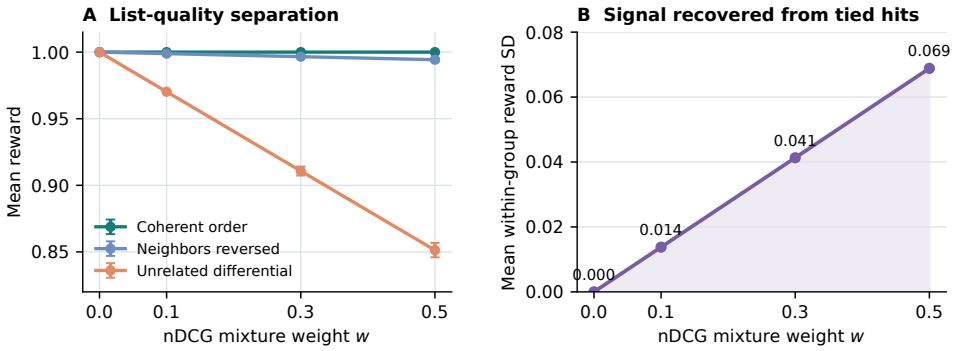

Figure 8: List-quality audit on 120 held-out diagnoses. Every list contains the exact diagnosis at rank 1. Panel A sweeps the nDCG weight with case-level bootstrap 95% intervals. Panel B reports within-case reward variation across coherent, reversed, and unrelated differentials.

K Adaptive retrieval-depth audit

Table 12 provides the numerical values underlying Figure 5. The frozen Qwen3.7-Plus judge returns the rank of the accepted reference diagnosis in each generated list, or zero when it is absent. We pool the four evaluation sources only for this behavioral analysis and retain every archived trajectory. Wilson intervals in the figure are computed independently within each depth stratum.

Table 12: ReAct diagnostic performance stratified by the number of retrieval calls selected by the policy. Values are percentages.

Calls	Cases	Hit@1	Hit@5	Hit@10
0	324	2.16	3.40	4.01
1	383	35.77	**54.31**	57.96
2	1,423	**38.65**	54.25	**59.31**
3	4,100	22.10	37.90	43.32

L LLM-as-a-Judge agreement and sensitivity audit

Evaluation scope. Table 1 and all headline comparisons use the deterministic disease normalizer; an LLM never assigns correctness in the primary evaluation. LLM-as-a-Judge is used only in auxiliary analyses of adaptive retrieval, where a generated disease description may denote the reference despite differing lexically from its canonical name. Its task is therefore narrow: adjudicating disease-identity equivalence, not assessing whether a diagnosis is clinically supported by the patient record. We audit this choice on identical archived Direct and ReAct outputs from seven benchmarks, giving 14 paired system-benchmark cells at each cutoff. Phenopackets is not omitted post hoc: it uses a different backbone and a dedicated three-hop structured-retrieval comparison (Table 2, Panel B), rather than the paired Direct/ReAct protocol in Panel A, and is consequently analyzed separately.

Blinded adjudication protocol. The fixed Qwen3.7-Plus judge configuration receives one reference disease and an anonymized ranked list. It is not shown the generating system, retrieval strategy, patient record, or benchmark. The judge returns the earliest rank at which a candidate denotes the same disease through an exact name, recognized synonym, abbreviation, or minor formatting variant; it returns zero when no such candidate exists. Thus, the judge cannot reward fluent explanations or method-specific traces. Its response is restricted to one integer under a 20-token output budget. API failures are retried and are never silently recoded as incorrect diagnoses. All 12,460 archived calls have a recorded judgment. The 286 empty model generations remain diagnostic failures rather than evaluator failures. Because the archived client did not set sampling parameters explicitly, provider defaults apply. We therefore make no claim of deterministic adjudication and evaluate sensitivity to the choice of evaluator instead.

Table 13: Macro comparison of two evaluators on identical Direct and ReAct outputs. Values are unweighted means (%) over the seven benchmarks in Table 2, Panel A.

Evaluator	Strategy	Hit@1	Hit@5	Hit@10
Deterministic	Direct	17.03	26.63	31.27
Deterministic	ReAct	16.49	27.51	33.53
Qwen3.7-Plus	Direct	23.54	36.46	41.97
Qwen3.7-Plus	ReAct	23.31	36.50	42.51

Agreement and sensitivity. Table 14 compares the 14 cell-level scores separately at each cutoff. The LLM Judge assigns consistently higher absolute scores, with a mean increase of 6.67-9.84 points, which is compatible with accepting variants missed by lexical normalization but may also include false-positive equivalences. The two score scales are therefore not interchangeable. Their within-cutoff rank correlations range from 0.785 to 0.881; a secondary pooled correlation over all 42 cell-cutoff measurements is 0.907, but we do not use the pooled value as primary evidence because score distributions differ across cutoffs. At the macro level, both evaluators associate ReAct with a small Hit@1 decrease and Hit@5/10 increases. At Hit@10, each favors ReAct on five of seven benchmarks, while their directions agree on five. MyGene2 and HMS reverse sign and are reported explicitly in Table 15. These discrepancies motivate retaining the deterministic evaluator for primary claims and restricting the LLM Judge to descriptive and depth-conditioned analyses.

Interpretation and limitations. The audit supports limited aggregate-level stability, while revealing non-negligible benchmark-level evaluator sensitivity. It is not a clinical validation of the LLM Judge: the deterministic matcher is a reproducible comparator, not a medical gold standard, and agreement between the two evaluators cannot establish that either resolves every synonym correctly. A blinded clinician-adjudicated sample, enriched for evaluator disagreements, would be required to estimate equivalence-level false-positive and false-negative rates. We therefore make no primary performance claim from LLM-judged scores. The reproducibility package records the complete Judge prompt, per-case integer outputs, retry logic, empty-generation handling, and both evaluator summaries, allowing this sensitivity analysis to be reproduced without regenerating predictions.

Table 14: Cross-evaluator sensitivity across 14 paired Direct/ReAct benchmark cells. ρ is Spearman correlation; MAE is the mean absolute difference, and shift is Qwen3.7-Plus minus the deterministic score, both in percentage points. Direction is the number of benchmarks for which the sign of ReAct minus Direct agrees.

Cutoff	ρ	MAE	LLM shift	Direction
Hit@1	0.881	7.67	+6.67	4/7
Hit@5	0.873	10.84	+9.41	6/7
Hit@10	0.785	11.76	+9.84	5/7

Table 15: Per-benchmark ReAct-minus-Direct change at Hit@10. Values are percentage points; directional disagreements are shown in bold.

Benchmark	Deterministic	Qwen3.7-Plus
MyGene2	**+7.6**	**−2.0**
RAMEDIS	−2.2	−7.1
MME	+2.5	+2.5
HMS	**−1.2**	**+1.1**
LIRICAL	+3.8	+2.1
RDS	+3.4	+3.8
RDC	+1.9	+3.4

M EFFICIENCY MEASUREMENT DETAILS

The archived 9B and 27B runs use eight data-parallel workers, tensor parallelism of one, greedy decoding, a 4,096-token context limit, and the same 6,249-record generation suite. The suite includes the main test records and auxiliary validation splits, so its size differs from Table 4. Per-shard generation times for 9B range from 128.4 to 137.2 seconds; the corresponding 27B range is 321.8 to 340.0 seconds. Node-level throughput is computed as the total number of cases divided by the maximum shard time, since evaluation completes when the slowest worker finishes. Model initialization, result merging, disease normalization, and metric computation are excluded.

Table 16: Computational profile. M is the number of sampled lists and T the number of ReAct turns.

Strategy	LLM generations	Local retrieval	External API
Direct	1	None	No
Static RAG	1	Once	No
Adaptive ReAct	T	Adaptive	No
Structured 3-hop	1	HPO + gene lookup	No
RRF fusion	M	Strategy-dependent	No

N TRAINING AND IMPLEMENTATION DETAILS

The 9B RL run uses two nodes with eight A800 80GB GPUs per node. The global training batch and generation batch are both 512, with 16 rollouts per prompt and a PPO minibatch size of 128. Prompt and response limits are 1,024 and 768 tokens. The actor learning rate is 5×10^{-7}; the reference-policy KL coefficient is 0.15. Policy-ratio clips are 0.20/0.28, and rollout/validation temperatures are 1.0/0.7. Zero-variance groups are resampled; training lasts at most 300 steps with checkpoints every ten steps, selected by held-out validation rather than training reward.

O SUPPLEMENTARY MODEL SCALE ABLATION

We are interested in investigating that if using a larger-scale LLM as base model can lead to better performance or not, and thus we isolate the effect of additional sampled diagnosis lists in a supplementary 128-case development-set experiment. This experiment uses fixed round-60, rank-64 LoRA checkpoints of Qwen3.8-27B and Qwen3.6-35B-A3B using tinker. These are separate checkpoints and a separate evaluation protocol from the main benchmark tables; the following values should not be compared directly with those tables. The cases are drawn from the 932-case development pool and are disjoint from an earlier 128-case pilot, but are not a new held-out test set. The validation-overlap limitations discussed above still apply. Scoring uses normalized exact matches to the gold diagnosis and its supplied aliases, without vocabulary snapping or an LLM judge.

For each case, we retain the first diagnosis from a greedy list and fill the remaining nine positions by reciprocal-rank voting over k additional lists. Each unique name contributes $1/r$ at its first position r in each sampled list; ties are broken lexicographically. No gold labels are used in aggregation. Additional lists use temperature 1.0, top-$p = 0.95$, top-$k = 50$, and a 1,024-token output limit, without retrieval or tools. All settings reuse the same recorded outputs: $k = 1, 2, 4$ use fixed prefixes of the eight samples. These intermediate budgets were examined post hoc; the $k = 8$ aggregation rule was fixed before this expanded evaluation. Top-1 is preserved by construction, at 28.91% for 27B and 28.12% for 35B.

At $k = 8$, Hit@10 increases by 11.72 percentage points for 27B and 9.38 points for 35B relative to greedy decoding, at approximately nine times the generated-token budget. For 35B, $k = 4$ and $k = 8$ have the same aggregate Hit@10 on this subset, although their correct cases differ; this does not establish general saturation. The experiment supports a coverage-generation-budget trade-off for this aggregation rule, not an improvement in model weights or first-choice accuracy. It also does not isolate the causal benefit of voting from that of extra sampling. As the 35B model has less active parameter size, the lower performance matches our expectation.

	Qwen3.8-27B		Qwen3.6-35B-A3B	
Extra lists k	Hit@10 (%)	Output tokens	Hit@10 (%)	Output tokens
0	34.38	173	32.81	179
1	36.72	346	32.81	354
2	39.06	519	36.72	529
4	39.84	864	42.19	877
8	46.09	1,556	42.19	1,575

Table 17: Sampling-budget ablation on 128 development cases with fixed model weights. Output tokens are mean totals per case across all $k+1$ generations; they exclude input tokens, retrieval, and runtime overhead.