

ORBIT: Traceable Rare-Disease Diagnosis via Orphanet-Grounded Retrieval and Bounded LLM Interaction

Songchi Zhou¹, Haichao Chen², Sheng Yu¹

¹Department of Statistics and Data Science, Tsinghua University

²Department of Ophthalmology, Peking Union Medical College Hospital

Correspondence: zhou-sc23@mails.tsinghua.edu.cn

Abstract

Rare-disease diagnosis requires converting heterogeneous patient descriptions into a ranked differential over thousands of disorders while keeping each hypothesis tied to verifiable evidence. Phenotype-based tools provide structured, interpretable comparisons but usually require normalized Human Phenotype Ontology (HPO) input. Large language models (LLMs) can read clinical narratives and propose broader hypotheses, yet their free-form outputs are difficult to ground in a disease catalog or combine with ontology-based evidence. We present ORBIT (Orphanet-grounded Retrieval and Bounded LLM Interaction for Traceable rare-disease diagnosis), a catalog-grounded diagnosis layer for both free-text and structured-HPO cases. ORBIT grounds open disease proposals to Orphanet, fuses them with exact phenotype and case-profile retrieval, scores only the resulting candidate pool with a probabilistic model, and reranks a fixed menu of evidence-bearing disease cards. The output records catalog identifiers, candidate provenance, phenotype support, and conflicts. Across RareArena and RareBench, ORBIT improves every reported metric over single-agent and medical-agent systems using the same backbone. Retrieval analysis, ablations, and expert review further identify where complementary candidate sources, iterative refinement, and constrained reranking affect performance.

1 Introduction

Rare diseases are individually uncommon but collectively affect hundreds of millions of people worldwide (Rehm, 2022; Haendel et al., 2020; Nguengang Wakap et al., 2020). For many patients, diagnosis remains a prolonged diagnostic odyssey: heterogeneous symptoms, incomplete records, limited specialist access, and overlapping disease presentations make it difficult to compare a patient against thousands of candidate disorders (Schieppati et al., 2008; Vandeborne et al., 2019; Benito-

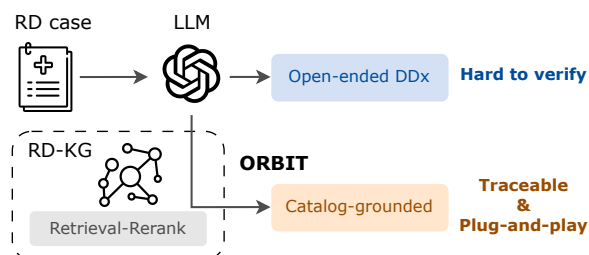


Figure 1: Open-ended LLM diagnosis versus catalog-grounded rare-disease reasoning. ORBIT retrieves and reranks catalog evidence, producing traceable catalog diagnoses rather than unconstrained free-form outputs.

Lozano et al., 2022; Glaubitz et al., 2025; Grier et al., 2018). AI-assisted differential diagnosis is therefore useful only if it can do more than name a plausible disease. In clinical or benchmark settings, the system needs to produce a ranked differential, ground each committed hypothesis to a curated disease identifier, and expose enough evidence for the result to be checked (Amorim et al., 2025; Germain et al., 2025; Groza et al., 2026).

Existing AI-assisted approaches address different parts of this requirement. The HPO provides a shared substrate for computable phenotyping, and phenotype-driven systems such as Phenomizer, PubCaseFinder, Exomiser, and LIRICAL compare patient HPO profiles with curated disease annotations to provide interpretable evidence trails (Robinson et al., 2008; Köhler et al., 2009; Fujiwara et al., 2018; Smedley et al., 2015; Robinson et al., 2020). However, they typically assume that the input has already been converted into a clean set of structured phenotype terms. Real cases are less tidy: evidence may be scattered across narratives, positives and negatives may be mixed, and the available phenotype set may be sparse or noisy.

LLM-based and agentic systems offer a complementary capability. They can read narratives, integrate dispersed findings, and propose broad differentials from latent medical knowledge (Chen

et al., 2024, 2026a,b; Zhao et al., 2026). Yet their outputs are often prompt-sensitive, free-form, and weakly tied to disease catalogs. Recent evaluations also show that LLMs do not uniformly dominate traditional rare-disease tools (Reese et al., 2026; Nguyen et al., 2026). This creates two coupled bottlenecks: candidate recall should benefit from flexible language understanding, while final decisions must remain grounded, auditable, and comparable across systems.

We introduce ORBIT, a framework for Orphanet-grounded rare-disease differential diagnosis. Rather than replacing an LLM or medical agent, ORBIT acts as a grounding and ranking layer around it. It can run directly on a patient case or accept disease-name proposals from an external agent and convert them into catalog identifiers. Internally, ORBIT normalizes case evidence into HPO terms, combines an open LLM proposal with exact HPO retrieval and case-profile match, applies phenotype scoring only to the candidate pool, and reranks fixed disease cards. Its interface accepts a case and optional candidate names, then returns catalog diagnoses with evidence traces. This boundary lets an agent call ORBIT without changing its high-level reasoning loop.

Figure 1 contrasts an open-ended LLM differential, whose free-form disease names are hard to verify against a fixed catalog, with ORBIT’s catalog-grounded output. ORBIT makes this output *traceable* by recording, for each final disease, its Orphanet identifier, the retrieval branches that proposed it, the HPO support or conflict used by the scorer, and the disease card shown to the reranker. These records do not establish clinical correctness, but they allow a reviewer to inspect the evidence and locate an error in retrieval, reasoning, or ontology mapping.

Our contributions are threefold. First, we formulate rare-disease diagnosis as bounded catalog-level reasoning that joins narrative hypotheses and phenotype retrieval in one Orphanet candidate state. Second, we introduce a controlled inference workflow that applies probabilistic HPO scoring within this state, validates iterative updates, and constrains final ranking to an evidence-bearing menu. Third, we evaluate both diagnostic performance and the behavior of the workflow through branch-coverage analysis, component ablations, expert trace review, and integration with an external medical agent.

2 Related Work

2.1 AI-Assisted Rare-Disease Diagnosis

Computer-assisted rare-disease diagnosis spans phenotype matching, variant-aware prioritization, patient-similarity retrieval, and LLM reasoning. HPO-driven tools score patient phenotypes against curated disease profiles using semantic similarity or likelihood ratios (Köhler et al., 2009; Fujiwara et al., 2018; Smedley et al., 2015; Robinson et al., 2020). Specialized LLMs and agent systems such as RareAgents and DeepRare (Chen et al., 2026b; Zhao et al., 2026) extend the input space to narratives and interactive reasoning. These approaches solve different parts of the task. Structured tools provide grounded evidence but depend on normalized input, while LLM systems read narratives but need catalog and ranking controls. ORBIT connects these roles: an LLM proposes diagnoses from the narrative, while Orphanet/HPO retrieval and scoring ground and order the candidate set. All stages use Orphanet identifiers and retain source and phenotype evidence.

2.2 Neuro-Symbolic and Retrieval-Augmented Reasoning

Retrieval-augmented generation conditions language models on external knowledge rather than only parametric memory (Lewis et al., 2020). Retrieve-then-rerank pipelines and rank aggregation also predate this work. Reciprocal rank fusion (RRF), for example, combines ranked lists without requiring calibrated scores or supervised fusion data (Cormack et al., 2009). Rare-disease LLM systems already use related forms of structured retrieval. RareBench constructs knowledge-graph-guided demonstrations and DeepRare combines disease normalization with knowledge and similar-case retrieval (Chen et al., 2024; Zhao et al., 2026).

Probabilistic ontology-based scoring also has a well-established lineage. BOQA integrates ontology structure and annotation frequencies within a Bayesian network while explicitly modeling false-positive and false-negative query noise (Bauer et al., 2012). Neuro-symbolic AI and ProbLog provide general machinery for combining neural representations with uncertain symbolic rules (Yu et al., 2023; Wan et al., 2024; Kimmig et al., 2011; De Raedt and Kimmig, 2015).

Building on these lines of work, ORBIT contributes a bounded catalog-level workflow that connects complementary proposal and retrieval routes

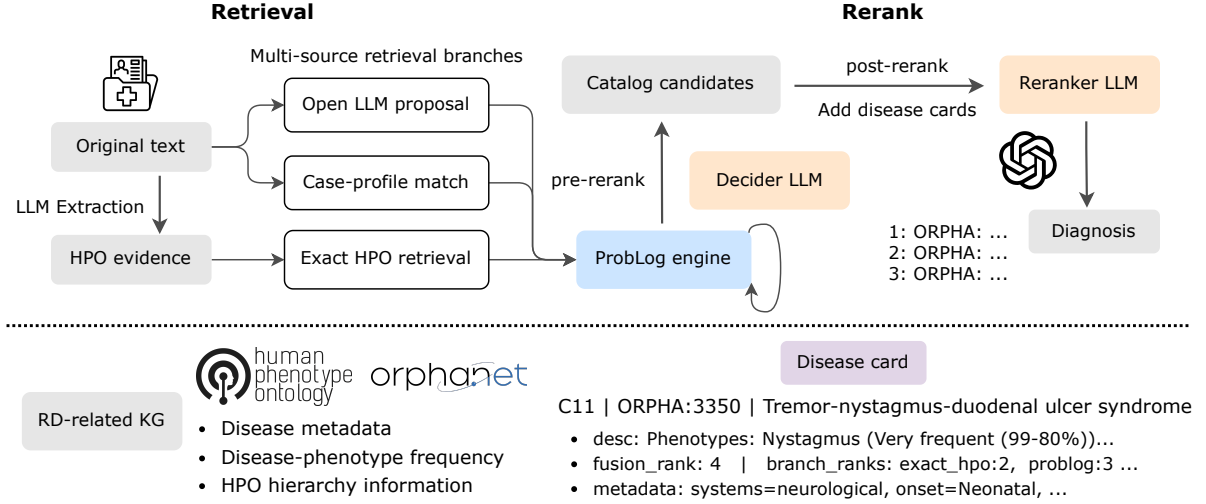


Figure 2: Modular ORBIT workflow: evidence normalization, three-branch catalog retrieval, candidate-restricted HPO scoring, and constrained reranking over fixed disease cards.

with candidate-restricted probabilistic scoring, validated refinement, and constrained reranking. The shared Orphanet interface retains source and phenotype evidence for each final candidate.

3 Methodology

Figure 2 summarizes the full workflow and the information passed between stages.

3.1 Task and Evidence State

For a case x , ORBIT returns a ranked differential over an Orphanet catalog $\mathcal{D}$. Orphanet supplies disease identifiers, canonical names, aliases, descriptions, and phenotype annotations. HPO supplies the phenotype vocabulary and hierarchy. An external agent may provide a case or candidate names, and ORBIT returns identifiers, source ranks, phenotype evidence, and a constrained ranking. The input may contain a narrative T , observed HPO terms P_0 , negated HPO terms N_0 , and metadata. After normalization, the active evidence state is:

$$E = (P, N, q^+, q^-), \quad P, N \subseteq \mathcal{H}, \quad (1)$$

where $\mathcal{H}$ is the HPO vocabulary, q_h^+ and q_h^- denote the reliability of positive and negative evidence for term h , respectively. Structured HPO inputs receive weight 1. For free-text cases, affirmative phenotype mentions are mapped by exact identifiers or top-1 embedding search over HPO names and synonyms. If mention m maps to HPO term h , $q_m \in [0, 1]$ is its normalizer confidence. The term-level positive reliability q_h is the noisy-or aggregation over

accepted mentions:

$$q_h = 1 - \prod_{m \in \mathcal{M}(h)} (1 - q_m), \quad (2)$$

where $\mathcal{M}(h)$ is the set of accepted mentions mapped to h , which is used as q_h^+ for $h \in P$. Narrative negation is not inferred automatically; only structured negated HPO input contributes to N and receives $q_h^- = 1$.

3.2 Grounded Candidate Construction

We represent each candidate source as an ordered list of Orphanet diseases. The open LLM proposal supplies broad hypotheses from narrative context but requires catalog grounding. Exact HPO retrieval provides direct, inspectable phenotype overlap but can miss diseases when HPO evidence is sparse. Case-profile match compares the complete case with disease profiles and can recover candidates with little exact overlap. We collect these three immutable source rankings in the retrieval branch map:

$$\mathcal{B}_{\text{ret}} = \{\text{llm} : L_{\text{llm}}, \text{ex} : L_{\text{ex}}, \text{prof} : L_{\text{prof}}\}. \quad (3)$$

Open LLM proposal. An LLM proposes an ordered differential in disease names. A proposed name n becomes a catalog candidate only if it maps to an Orphanet canonical name or alias, or if embedding search over disease names and aliases is above a grounding threshold. Let e_n be the embedding of n , e_d the embedding of catalog disease d , sim cosine similarity, and τ_g the threshold. We

define:

$$d^*(n) = \arg \max_{d \in \mathcal{D}} \text{sim}(e_n, e_d),$$

$$g(n) = \begin{cases} d^*(n), & \text{sim}(e_n, e_{d^*(n)}) \geq \tau_g, \\ \emptyset, & \text{otherwise,} \end{cases} \quad (4)$$

where $g(n)$ is the committed catalog identifier. The grounded names, in their proposal order after deduplication, form L_{llm} .

Exact HPO retrieval. This branch ranks diseases by direct overlap with the observed HPO set P . Let $\text{df}(h) = |\{d \in \mathcal{D} : h \in H_d\}|$ be the number of diseases annotated with HPO term h . The smoothed inverse-document-frequency (IDF) weight is:

$$I(h) = 1 + \log \frac{|\mathcal{D}| + 1}{\text{df}(h) + 1}, \quad (5)$$

so rare, more discriminative HPO terms receive larger weights while ubiquitous terms remain down-weighted. Let H_d be the catalog phenotype profile of disease d , and let $O_d = P \cap H_d$ be the exact matched terms. For any term set S , define its total and Euclidean IDF mass as $u_S = \sum_{h \in S} I(h)$ and $n_S = (\sum_{h \in S} I(h)^2)^{1/2}$. For disease d ,

$$m_d = \sum_{h \in O_d} I(h), \quad z_d = \sum_{h \in O_d} I(h)^2, \quad (6)$$

$$a_d = \frac{m_d}{u_P}, \quad b_d = \frac{m_d}{u_{H_d}}, \quad c_d = \frac{z_d}{n_P n_{H_d}}.$$

Here a_d is case-evidence coverage, b_d is disease-profile concentration, and c_d is IDF-weighted cosine overlap. The branch score is:

$$s_{\text{ex}}(d) = \left(c_d \frac{2a_d b_d}{a_d + b_d} \right)^{1/2}, \quad (7)$$

with zero score for empty-overlap or zero-denominator cases. Sorting by s_{ex} gives L_{ex} .

Case-profile match. The profile branch embeds a query built from the narrative, age context, and normalized HPO cues. Disease profiles are embedded from catalog names, aliases, descriptions, onset, inheritance, affected systems, and frequent HPO phenotypes. With case embedding e_x and disease-profile embedding v_d , the branch score is:

$$s_{\text{prof}}(d) = \text{sim}(e_x, v_d). \quad (8)$$

Let B_{prof} be the retrieval budget. We first keep the top B_{prof} diseases by s_{prof} , then return an adaptive prefix L_{prof} using the similarity-gap cutoff defined in Appendix B.1; if no clear gap is detected, the full budgeted list is retained.

3.3 Catalog Candidate Fusion

The three grounded branches are fused before probabilistic scoring. Let $S(d) = \{b : d \in L_b\}$ be the set of branches that returned disease d , and let $r_b(d)$ be its one-indexed rank in branch b . Non-empty branches are combined with equal-weight RRF using $K = 60$:

$$s_{\text{rrf}}(d) = \sum_{b \in S(d)} \frac{1}{K + r_b(d)}. \quad (9)$$

The seed list $L_{\text{pool}}^{(0)}$ is the deduplicated union of branch outputs sorted by Eq. (9), and $\mathcal{C}_0$ is the corresponding set of diseases. Ties are broken by more supporting branches $|S(d)|$, then by the best branch position:

$$r_{\min}(d) = \min_{b \in S(d)} r_b(d), \quad (10)$$

and finally by earliest appearance under a fixed branch order.

3.4 Candidate-Restricted HPO Scoring

After retrieval fusion, ORBIT uses the normalized HPO evidence to compare a bounded set of grounded hypotheses. At iteration t , $L_{\text{pool}}^{(t)}$ is the ordered pool and $\mathcal{C}_t$ is its underlying disease set. We implement the scorer with ProbLog (Kimmig et al., 2011): each iteration instantiates a probabilistic program over $\mathcal{C}_t$, rather than asking an LLM to assign unconstrained disease probabilities. The program contains a mutually exclusive diagnosis choice, disease-to-phenotype links from catalog frequencies, and noisy observation variables for the positive and negative evidence.

At iteration t , the validated refinements Π_t induce an active evidence set $E_t = (P_t, N_t, q^+, q^-)$ from the normalized evidence E (Appendix B.2). We instantiate the ProbLog scorer only when the active positive evidence meets the threshold $|P_t| \geq \tau_{\text{PL}}$; otherwise the posterior rank signal is omitted.

For a candidate disease d and active HPO term h , let y_{dh}^s , for $s \in \{+, -\}$, denote the polarity-specific phenotype probability obtained from HPO ancestor closure and the per-case adaptive leak probability in Eq. (30). The leak term assigns probability to phenotype evidence left unexplained by catalog associations in the seed pool. The evidence reliabilities q_h^+ and q_h^- define local likelihoods $\ell_h^+(d)$ and $\ell_h^-(d)$ for observing or negating term h under diagnosis d . ProbLog conditions on

the active evidence and ranks candidates by the posterior:

$$\Pr(d \mid \mathcal{S}_t) \propto \pi_d \prod_{h \in P_t} \ell_h^+(d) \prod_{h \in N_t} \ell_h^-(d), \quad (11)$$

where π_d is the candidate prior. The construction of π_d , y_{dh}^s , and the noisy evidence likelihoods is given in Appendix B.3.

The posterior-ranked catalog list is denoted $\mathcal{R}_{\text{PL}}^{(t)}$ and is passed downstream as an additional ProbLog rank signal. Between scoring rounds, the bounded decider may stop, add groundable disease names, or request validated refinements. These operations do not add new clinical observations. Instead, they update the candidate set, scoring constraints, or the active subset of the existing evidence used by ProbLog.

3.5 Constrained Final Reranking

After the last refinement iteration t^* , ORBIT builds a fixed catalog menu for the final LLM reranker. We keep three objects separate. The retrieval branch map $\mathcal{B}_{\text{ret}}$ preserves the original source rankings. The ordered state $L_{\text{pool}}^{(t^*)}$ is initialized by RRF over those branches and then extended, without reordering existing entries, by validated catalog additions. Its underlying set $\mathcal{C}_{t^*}$ defines which diseases are eligible for the final menu. The optional map $\mathcal{B}_{\text{PL}}^{(t^*)}$ contains the final ProbLog ranking:

$$\mathcal{B}_{\text{PL}}^{(t^*)} = \begin{cases} \{\text{pl} : \mathcal{R}_{\text{PL}}^{(t^*)}\}, & \text{if ProbLog is used,} \\ \emptyset, & \text{otherwise,} \end{cases} \quad (12)$$

The rank lists supplied to the menu-building fusion are:

$$\mathcal{B}_{\text{menu}}^{(t^*)} = \mathcal{B}_{\text{ret}} \cup \mathcal{B}_{\text{PL}}^{(t^*)}. \quad (13)$$

We compute equal-weight RRF over $\mathcal{B}_{\text{menu}}^{(t^*)}$ for diseases in $\mathcal{C}_{t^*}$ and retain the first M diseases:

$$\mathcal{M}_{\text{cat}} = \text{Top}_M \left(\text{RRF}(\mathcal{B}_{\text{menu}}^{(t^*)}; \mathcal{C}_{t^*}) \right). \quad (14)$$

Candidates added during refinement can belong to the domain even if they do not occur in a retained branch list; equal RRF ties preserve their order in $L_{\text{pool}}^{(t^*)}$. Each selected disease becomes a compact card containing its catalog identifier, name, aliases, description, branch ranks, and ProbLog information when available. In the reported main configuration, unmapped open names are excluded,

so the final LLM must output only identifiers from this fixed catalog menu:

$$\hat{R} = (d_1, \dots, d_K), \quad d_i \in \mathcal{M}_{\text{cat}}, \quad d_i \neq d_j. \quad (15)$$

Invalid or duplicated identifiers are removed, and missing slots are filled from the fused menu order.

4 Experiments

4.1 Experimental Setup

Datasets. The experiments incorporate biomedical ontology knowledge from Orphanet,¹ and the Human Phenotype Ontology (HPO) (Robinson et al., 2008). We evaluate on RareArena (Chen et al., 2026a) and RareBench (Chen et al., 2024). RareArena contributes 8,562 free-text cases from its released rare-disease screening test split, and RareBench contributes 1,197 structured-HPO cases from its released differential-diagnosis set. In both settings, Orphanet defines the disease catalog and disease metadata, while HPO provides the phenotype vocabulary, hierarchy, and disease-phenotype annotations used for normalization, retrieval, and scoring. Release versions and data-handling details are reported in Appendix A.1.

Baselines. We compare our method with three categories of baselines:

1. **General and medical LLMs.** We evaluate general and medical LLMs under a zero-shot chain-of-thought (CoT) setting (Wei et al., 2022), including DeepSeek-V3.2 (Liu et al., 2025) and Baichuan-M2 (Dou et al., 2025).
2. **Medical agent frameworks.** We compare with representative medical agent frameworks, including MedAgents (Tang et al., 2024) and MDAgents (Kim et al., 2024). For these agent-based baselines, we use Qwen3-30B-A3B-Instruct (Yang et al., 2025) and gpt-oss-20b (Agarwal et al., 2025) as the backbone agent models.
3. **Phenotype-based methods.** We further compare with phenotype-based rare disease diagnosis methods, including PhenoBrain (Mao et al., 2025), Exomiser (Smedley et al., 2015), and LIRICAL (Robinson et al., 2020). Since these methods take structured phenotypic representations as input, we only conduct comparisons with them on RareBench, where pa-

¹<https://www.orpha.net/>

Method	RareArena					RareBench				
	H@1	H@3	H@5	H@10	MRR@10	H@1	H@3	H@5	H@10	MRR@10
<i>General & Medical LLMs (zero-shot CoT)</i>										
DeepSeek-V3.2	0.1998	0.3159	0.3628	0.4167	0.3279	0.3183	0.4545	0.5372	0.6023	0.4043
Baichuan-M2	0.1550	0.2372	0.2692	0.3045	0.2664	0.3175	0.4503	0.5121	0.5831	0.4001
<i>Qwen3-30b-A3b (Medical agent framework)</i>										
Single-Agent	0.1177	0.2084	0.2483	0.2951	0.2241	0.2398	0.4052	0.4536	0.5113	0.3363
MedAgents	0.1701	0.2517	0.2867	0.3336	0.2602	0.2339	0.3434	0.4160	0.5063	0.3103
MDAgents	0.1507	0.2373	0.2728	0.3135	0.2477	0.2289	0.3409	0.4185	0.4996	0.3056
ORBIT	0.1754	0.2849	0.3326	0.3915	0.2803	0.2473	0.4144	0.5071	0.5848	0.3537
<i>gpt-oss-20b (Medical agent framework)</i>										
Single-Agent	0.1662	0.2446	0.2782	0.3078	0.2612	0.2918	0.4323	0.4833	0.5284	0.3729
MedAgents	0.2049	0.2847	0.3236	0.3559	0.2926	0.3002	0.4415	0.5184	0.5694	0.3881
MDAgents	0.1718	0.2502	0.2840	0.3191	0.2694	0.3161	0.4415	0.4983	0.5460	0.3923
ORBIT	0.2186	0.3306	0.3760	0.4321	0.3296	0.3383	0.4820	0.5631	0.6466	0.4331

Table 1: Performance on RareArena and RareBench. H@k denotes Hit@k under equivalent-diagnosis matching. RareArena uses score-2 matches.

tient information is provided in an HPO-based format.

RareAgents (Chen et al., 2026b) is omitted because its implementation and full coordination prompts are unavailable. We do not place the published DeepRare result (Zhao et al., 2026) in the main table because its similar-case bank overlaps RareArena and the covered case IDs are undisclosed. Instead, Section 4.6 reports a controlled reproduction with that module disabled and tests ORBIT as an external DeepRare tool.

Evaluation metrics. Following the source benchmarks, a matching judge compares each predicted disease name with the reference diagnosis without seeing the clinical case. RareArena distinguishes an equivalent or synonymous disease-level match (score 2), a listed less-specific hierarchical match (score 1), and a non-match (score 0). In our experiments, the main RareArena metrics count score-2 matches only. For both datasets, Hit@ k records whether a correct match occurs in the first k positions. MRR@10 is the reciprocal rank of the first match and is zero when none appears. We use DeepSeek-V3.2 as the judge model. Cross-judge results are reported in Appendix D.

4.2 Differential diagnosis

As shown in Table 1, ORBIT consistently improves the diagnostic performance across both RareArena and RareBench. Under each backbone model, ORBIT outperforms the corresponding Single-Agent, MedAgents, and MDAgents base-

	H@1	H@3	H@5	H@10	MRR@10
<i>Phenotype-based methods</i>					
PhenoBrain	0.2834	0.4590	0.5293	0.6321	0.3901
Exomiser	0.2266	0.3913	0.5059	0.6104	0.3393
LIRICAL	0.2375	0.3921	0.4666	0.5460	0.3329
ORBIT	0.3383	0.4820	0.5631	0.6466	0.4331

Table 2: Comparison between ORBIT and phenotype-based methods on RareBench. The result of ORBIT is based on the gpt-oss-20b backbone.

lines on all evaluation metrics, showing that the proposed method brings consistent gains over existing medical agent frameworks. When using gpt-oss-20b as the backbone, ORBIT achieves the best overall results among all methods. In particular, it surpasses the strongest zero-shot CoT baseline, DeepSeek-V3.2, on both benchmarks, improving H@10 from 0.4167 to 0.4321 on RareArena and from 0.6023 to 0.6466 on RareBench. ORBIT also achieves the highest MRR@10 scores, with 0.3296 on RareArena and 0.4331 on RareBench, indicating better ranking quality in differential diagnosis. These results demonstrate that ORBIT can effectively enhance both top-ranked diagnosis accuracy and broader candidate disease ranking across unstructured free-text and structured HPO-based settings.

On RareBench, Table 2 compares the gpt-oss-20b ORBIT configuration with phenotype-based tools. Notably, ORBIT achieves the best performance across all metrics. Compared with the strongest phenotype-based baseline, PhenoBrain,

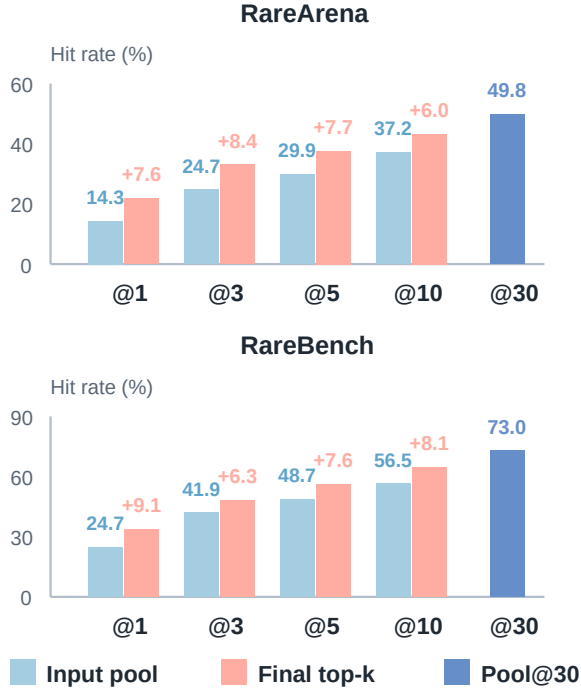


Figure 3: Pool-to-rank analysis on RareArena and RareBench. The fused reranker input pool provides the initial Hit@ k order, while final reranking improves early ranks. Annotations denote absolute percentage-point gains. Pool@30 reports whether the gold diagnosis appears in the top-30 pool before reranking.

ORBIT improves H@1 from 0.2834 to 0.3383, H@10 from 0.6321 to 0.6466, and MRR@10 from 0.3901 to 0.4331. The improvement in H@1 and MRR@10 suggests that ORBIT is better at ranking the correct diagnosis near the top, while the consistent gains in H@10 indicate stronger differential diagnosis coverage. Thus, ORBIT can effectively leverage structured phenotype information and outperform specialized phenotype-based methods in HPO-based rare disease diagnosis.

4.3 Retrieval and Reranking Diagnostics

The diagnostic analysis separates candidate coverage from final ordering and measures overlap between the three retrieval branches.

Figure 3 compares the fused input order with the final disease-card ranking. On RareArena, reranking raises Hit@1, Hit@3, Hit@5, and Hit@10 by 7.6, 8.4, 7.7, and 6.0 percentage points. The RareBench gains are 9.1, 6.3, 7.6, and 8.1 points. Pool@30 is 49.8% on RareArena and 73.0% on RareBench, which leaves room for both candidate recall and within-pool ordering to improve.

Tables 3 and 4 explain the benefit of fusion. All-branch Jaccard overlap is 0.013 on RareArena

Branch pair	RareArena	RareBench
LLM prop. – HPO ret.	0.025	0.035
LLM prop. – Case profile	0.079	0.069
HPO ret. – Case profile	0.085	0.085
All three branches	0.013	0.017

Table 3: Mean Jaccard overlap between retrieval branches over Orphanet identifiers before fusion. LLM prop.: open LLM proposal; HPO ret.: exact HPO retrieval; Case profile: case-profile match. All three: the intersection-over-union across the three sets.

Branch	RareArena		RareBench	
	Mean	Gold (%)	Mean	Gold (%)
Open LLM proposal	13.1	33.1	17.4	48.7
Exact HPO retrieval	18.4	18.3	19.8	43.4
Case-profile match	8.7	32.6	8.8	50.8

Table 4: Branch coverage before fusion. Mean is the number of unique grounded candidates returned per case. Gold is the percentage of cases whose full branch output contains a correct match.

and 0.017 on RareBench, and every pairwise value is below 0.09. The branches therefore cover different catalog regions. The open LLM proposal covers 33.1% and 48.7% of cases, exact HPO retrieval covers 18.3% and 43.4%, and case-profile match covers 32.6% and 50.8% on RareArena and RareBench, respectively. The highest-coverage branch differs by dataset: the open LLM proposal leads on RareArena, whereas case-profile match leads on RareBench.

4.4 Ablations

We fix both subsets before running any ablation: 495 RareArena cases and 403 RareBench cases. Every variant uses gpt-oss-20b, identical prompts and decoding settings, and the same case IDs. Appendix A.2 gives the subset construction.

Removing the final reranker produces the largest drop in Table 5: 0.1208 on RareArena and 0.0839 on RareBench. Case-profile match and the open LLM proposal also have effects well above the observed judge variation. Removing exact HPO retrieval lowers Hit@10 by 0.0246 and 0.0362, while removing ProbLog inference lowers it by 0.0170 and 0.0174. Disabling the refinement loop while retaining retrieval, the initial ProbLog pass, and final reranking lowers Hit@10 by 0.0238 and 0.0233. These results distinguish the contribution of direct phenotype retrieval from the subsequent probabilis-

	RareArena	RareBench
ORBIT	0.5406 $\pm$ 0.0023	0.6670 $\pm$ 0.0103
<i>Removing individual modules</i>		
w/o Exact HPO retrieval	0.5160 $\pm$ 0.0056	0.6308 $\pm$ 0.0048
w/o Case-profile match	0.4347 $\pm$ 0.0039	0.6000 $\pm$ 0.0088
w/o Open LLM proposal	0.4412 $\pm$ 0.0079	0.6248 $\pm$ 0.0081
w/o ProbLog inference	0.5236 $\pm$ 0.0056	0.6496 $\pm$ 0.0037
w/o Final rerank	0.4198 $\pm$ 0.0048	0.5831 $\pm$ 0.0093
w/o Iterative refinement	0.5168 $\pm$ 0.0036	0.6437 $\pm$ 0.0051
<i>Adding auxiliary module</i>		
+ Unmapped LLM proposal	0.5103 $\pm$ 0.0051	0.6655 $\pm$ 0.0069

Table 5: Hit@10 on fixed ablation subsets, reported as mean $\pm$ std across five judge runs.

tic scoring and iterative refinement. Passing unmapped LLM proposal directly to the reranker lowers Hit@10 to 0.5103 on RareArena and 0.6655 on RareBench. Based on this result, we use the catalog-only setting in the reported configuration.

4.5 Trace Audit

We conduct two expert analyses with distinct random samples. The first trace audit covers 200 cases: 100 correctly answered cases and 100 incorrectly answered cases. For each correct case, clinical experts score the model-generated reasoning on reasoning correctness (0–3), evidence comprehensiveness (0–3), and case-specific grounding (0–2). For a maximum of eight points, a total above 5 is treated as clinically acceptable.

In the correct-case sample, 87% score above 5 of 8; 85% receive at least 2 for reasoning correctness, 81% receive at least 2 for evidence comprehensiveness, and 99% demonstrate appropriate use of case-specific evidence. For the incorrectly answered cases, four major sources of error were identified: retrieval failure, accounting for 60% of cases; reasoning failure, accounting for 23%, such as overemphasizing complications while overlooking disease-specific manifestations; ontology-granularity errors, accounting for 10%, in which the predicted diagnosis was clinically related to the reference diagnosis but expressed at an inappropriate level of specificity; and diagnostic-level mismatch, accounting for 7%, in which the model identified a superficial clinical syndrome rather than the underlying disease.

Separately, experts assess the final reranking rationale for 50 randomly selected cases using the same three dimensions and eight-point total. A rationale is treated as clinically acceptable when its

Method	RareArena	RareBench
Single Agent	0.3345 $\pm$ 0.0076	0.5129 $\pm$ 0.0099
ORBIT	0.5406 $\pm$ 0.0023	0.6670 $\pm$ 0.0103
DeepRare	0.3370 $\pm$ 0.0122	0.5648 $\pm$ 0.0080
DeepRare + ORBIT tool	0.5168 $\pm$ 0.0071	0.6541 $\pm$ 0.0071

Table 6: Controlled DeepRare integration with gpt-oss-20b as the base agent. Hit@10 is reported as mean $\pm$ std across five judge runs on the fixed subsets.

	RareArena	RareBench
MedAgents	13.8	17.5
MDAgents	39.2	32.4
ORBIT	3.4	3.5

Table 7: Average number of LLM calls per case on the ablation subset.

total exceeds 5. Overall, 34 of the 50 evaluated cases (68%) met this threshold, suggesting that the reranking model generally bases its decisions on clinically relevant evidence and provides reasonably sound and sufficiently grounded justifications.

Neither analysis establishes clinical safety. Together, they test whether the recorded reasoning exposes case-specific evidence and a locatable failure stage.

4.6 Agent integration

Our DeepRare reproduction disables similar-case retrieval to avoid undisclosed RareArena overlap and disables PubCaseFinder because its public API rate limit prevents a reproducible large-scale run. All configurations in this controlled comparison use gpt-oss-20b as the base agent. Adding ORBIT as an external tool, without changing DeepRare’s high-level workflow, raises Hit@10 by 17.98 percentage points on RareArena and 8.93 points on RareBench. The result supports interface compatibility of ORBIT with existing agentic systems.

4.7 Cost analysis on LLM calls

On the ablation subsets, ORBIT uses 3.4 LLM calls per RareArena case and 3.5 per RareBench case (Table 7). MedAgents uses 13.8 and 17.5, while MDAgents uses 39.2 and 32.4. Each ORBIT case receives a proposal call and a final reranking call. Free-text cases also require HPO extraction, and the decider runs only when the current symbolic state has an actionable update. Retrieval and disease grounding take about 1 second per case on the same subsets. ProbLog takes 0.5 seconds on RareArena and 1.7 seconds on RareBench, while

evidence assembly and logging take about 1 second. These non-generative stages add a few seconds per case and do not dominate inference time.

5 Conclusion

We presented ORBIT, a plug-and-play Orphanet-grounded framework for rare-disease differential diagnosis. By separating open-ended hypothesis generation, catalog grounding, candidate-restricted HPO scoring, and constrained reranking, ORBIT can run as a standalone system or as a grounding layer inside existing agentic workflows. Experiments on RareArena and RareBench show consistent gains over single-agent, multi-agent, and phenotype-based baselines, while process analyses confirm that complementary retrieval and final reranking are both important. These results support bounded LLM interaction with catalog-grounded evidence as a practical direction for rare-disease decision support.

Limitations

ORBIT depends on the coverage and quality of Orphanet and HPO annotations. Diseases with incomplete phenotype profiles, ambiguous aliases, sparse frequency labels, or weak links between narrative findings and HPO terms may be under-ranked. The system also relies on embedding thresholds and LLM-generated extraction, proposal, and reranking outputs, which may vary across model backbones, prompt formats, and languages.

Our evaluation is benchmark-based and uses an LLM-as-judge matching protocol. This may not fully capture clinical equivalence, comorbid presentations, longitudinal disease evolution, or the way clinicians integrate genetic, laboratory, imaging, and family-history evidence. Although ORBIT is designed to plug into existing agents, we do not exhaustively test all possible orchestration strategies. ORBIT is intended for differential-diagnosis support rather than autonomous diagnosis. Prospective clinician-in-the-loop validation, local calibration, and integration with additional clinical modalities are needed before deployment.

Ethical considerations

Rare-disease diagnosis is high-stakes. ORBIT should be used only as assistive decision support under qualified clinical supervision, and its ranked outputs should not replace clinician judgment, confirmatory testing, or patient-specific evaluation. Be-

cause the system processes sensitive clinical narratives and phenotype data, deployment must follow applicable privacy, consent, logging, access-control, and data-governance requirements.

The method may reflect biases in curated catalogs, benchmark composition, and LLM training data, potentially affecting underrepresented populations or poorly annotated disorders. We emphasize catalog identifiers, evidence summaries, and deterministic backfill to support review, but users should still audit failure cases, monitor subgroup performance, and communicate uncertainty when predictions are used in clinical or research settings.

References

- Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, et al. 2025. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*.
- Ana MB Amorim, Urszula Orzeł, Ana B Caniceiro, Nícia Rosário-Ferreira, and Irina S Moreira. 2025. Artificial intelligence in rare diseases: toward clinical impact. *Trends in Pharmacological Sciences*.
- Sebastian Bauer, Sebastian Köhler, Marcel H Schulz, and Peter N Robinson. 2012. Bayesian ontology querying for accurate and noise-tolerant semantic searches. *Bioinformatics*, 28(19):2502–2508.
- Juan Benito-Lozano, Blanca López-Villalba, Greta Arias-Merino, Manuel Posada De la Paz, and Verónica Alonso-Ferreira. 2022. Diagnostic delay in rare diseases: data from the spanish rare diseases patient registry. *Orphanet Journal of Rare Diseases*, 17(1):418.
- Haichao Chen, Zhengyun Zhao, Songchi Zhou, Shikai Hu, Jinyuan Wang, Ye Jin, Xianghong Jin, Yih Chung Tham, Xiaofei Wang, Weizhi Ma, et al. 2026a. Rarearena: a comprehensive benchmark dataset unveiling the potential of large language models in rare disease diagnosis. *The Lancet Digital Health*, 8(2).
- Xuanzhong Chen, Ye Jin, Xiaohao Mao, Lun Wang, Shuyang Zhang, and Ting Chen. 2026b. Rareagents: Autonomous multi-disciplinary team for rare disease diagnosis and treatment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 101–109.
- 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 the 30th ACM SIGKDD conference on knowledge discovery and data mining*, pages 4850–4861.
- Gordon V Cormack, Charles LA Clarke, and Stefan Buettcher. 2009. Reciprocal rank fusion outperforms

- condorcet and individual rank learning methods. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 758–759.
- Luc De Raedt and Angelika Kimmig. 2015. Probabilistic (logic) programming concepts. *Machine Learning*, 100(1):5–47.
- Chengfeng Dou, Chong Liu, Fan Yang, Fei Li, Jiyuan Jia, Mingyang Chen, Qiang Ju, Shuai Wang, Shunya Dang, Tianpeng Li, et al. 2025. Baichuan-m2: Scaling medical capability with large verifier system. *arXiv preprint arXiv:2509.02208*.
- 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(3):389–399.
- Dominique P Germain, David Gruson, Marie Malcles, and Nicolas Garcelon. 2025. Applying artificial intelligence to rare diseases: a literature review highlighting lessons from fabry disease. *Orphanet Journal of Rare Diseases*, 20(1):186.
- Rick Glaubitz, Luise Heinrich, Falko Tesch, Martin Seifert, Katrin Christiane Reber, Ursula Marschall, Jochen Schmitt, and Gabriele Müller. 2025. The cost of the diagnostic odyssey of patients with suspected rare diseases. *Orphanet Journal of Rare Diseases*, 20(1):222.
- Johnston Grier, Michio Hirano, Amel Karaa, Emma Shepard, and John LP Thompson. 2018. Diagnostic odyssey of patients with mitochondrial disease: results of a survey. *Neurology: Genetics*, 4(2):e230.
- Tudor Groza, Gareth Baynam, and Saumya Shekhar Jamuar. 2026. Reimagining care of people living with rare diseases with artificial intelligence. *Plos Medicine*, 23(2):e1004966.
- Melissa Haendel, Nicole Vasilevsky, Deepak Unni, Cristian Bologna, Nomi Harris, Heidi Rehm, Ada Hamosh, Gareth Baynam, Tudor Groza, Julie McMurphy, et al. 2020. How many rare diseases are there? *Nature reviews drug discovery*, 19(2):77–78.
- Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik S Chan, Xuhai Xu, Daniel McDuff, Hyeonhoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, and Hae W Park. 2024. Mdagents: An adaptive collaboration of llms for medical decision-making. *Advances in Neural Information Processing Systems*, 37:79410–79452.
- Angelika Kimmig, Bart Demoen, Luc De Raedt, Victor Santos Costa, and Ricardo Rocha. 2011. On the implementation of the probabilistic logic programming language problog. *Theory and Practice of Logic Programming*, 11(2-3):235–262.
- Sebastian Köhler, Marcel H Schulz, Peter Krawitz, Sebastian Bauer, Sandra Dölken, Claus E Ott, Christine Mundlos, Denise Horn, Stefan Mundlos, and Peter N Robinson. 2009. Clinical diagnostics in human genetics with semantic similarity searches in ontologies. *The American Journal of Human Genetics*, 85(4):457–464.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.
- Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, et al. 2025. Deepseek-v3. 2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556*.
- Xiaohao Mao, Yu Huang, Ye Jin, Lun Wang, Xuanzhong Chen, Honghong Liu, Xinglin Yang, Haopeng Xu, Xiaodong Luan, Ying Xiao, et al. 2025. A phenotype-based ai pipeline outperforms human experts in differentially diagnosing rare diseases using ehers. *NPJ Digital Medicine*, 8(1):68.
- Stéphanie Nguengang Wakap, Deborah M Lambert, Annie Olry, Charlotte Rodwell, Charlotte Gueydan, Valérie Lanneau, Daniel Murphy, Yann Le Cam, and Ana Rath. 2020. Estimating cumulative point prevalence of rare diseases: analysis of the orphanet database. *European journal of human genetics*, 28(2):165–173.
- Minh-Ha Nguyen, Chih-Ting Yang, Thomas A Cassini, Fan Ma, Rizwan Hamid, Lisa Bastarache, Josh F Peterson, Hua Xu, Lingyao Li, Siyuan Ma, et al. 2026. Diagnostic accuracy of large language models for rare diseases: A systematic review and meta-analysis. *medRxiv*.
- Justin T Reese, Leonardo Chimirri, Yasemin Bridges, Daniel Danis, J Harry Caufield, Michael A Gargano, Carlo Kroll, Andrew Schmeder, Fengchen Liu, Kyran Wissink, et al. 2026. Systematic benchmarking demonstrates large language models have not reached the diagnostic accuracy of traditional rare-disease decision support tools. *European Journal of Human Genetics*, pages 1–7.
- Heidi L Rehm. 2022. Time to make rare disease diagnosis accessible to all. *Nature Medicine*, 28(2):241–242.
- Peter N Robinson, Sebastian Köhler, Sebastian Bauer, Dominik Seelow, Denise Horn, and Stefan Mundlos. 2008. The human phenotype ontology: a tool for annotating and analyzing human hereditary disease. *The American Journal of Human Genetics*, 83(5):610–615.
- 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. 2020. Interpretable clinical genomics with a likelihood ratio paradigm. *The American Journal of Human Genetics*, 107(3):403–417.
- Arrigo Schieppati, Jan-Inge Henter, Erica Daina, and Anita Aperia. 2008. Why rare diseases are an important medical and social issue. *The Lancet*, 371(9629):2039–2041.
- 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. 2015. Next-generation diagnostics and disease-gene discovery with the exomiser. *Nature protocols*, 10(12):2004–2015.
- Xiangru Tang, Anni Zou, Zhuosheng Zhang, Ziming Li, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. 2024. Medagents: Large language models as collaborators for zero-shot medical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 599–621.
- Liese Vandeborne, Eline van Overbeeke, Marc Doods, Birgit De Beleyr, and Isabelle Huys. 2019. Information needs of physicians regarding the diagnosis of rare diseases: a questionnaire-based study in belgium. *Orphanet journal of rare diseases*, 14(1):99.
- Zishen Wan, Che-Kai Liu, Hanchen Yang, Chaojian Li, Haoran You, Yonggan Fu, Cheng Wan, Tushar Krishna, Yingyan Lin, and Arijit Raychowdhury. 2024. Towards cognitive ai systems: a survey and prospective on neuro-symbolic ai. *arXiv preprint arXiv:2401.01040*.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.
- Dongran Yu, Bo Yang, Dayou Liu, Hui Wang, and Shirui Pan. 2023. A survey on neural-symbolic learning systems. *Neural Networks*, 166:105–126.
- Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, et al. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.
- Weike Zhao, Chaoyi Wu, Yanjie Fan, Pengcheng Qiu, Xiaoman Zhang, Yuze Sun, Xiao Zhou, Shuju Zhang, Yu Peng, Yanfeng Wang, et al. 2026. An agentic system for rare disease diagnosis with traceable reasoning. *Nature*, pages 1–10.

A Experimental Details

A.1 Data Releases and Reproducibility

All disease profiles, phenotype annotations, and retrieval indices use the HPO 2024-07-01 release and Orphanet metadata version 1.3.42 / 4.1.8 [2025-03-03]. The main evaluation uses the complete released RareArena RDS test split and RareBench differential-diagnosis set without additional case filtering. RareArena is available at <https://github.com/zhao-zy15/RareArena>, and RareBench is available at <https://github.com/chenxz1111/RareBench>. Qwen3-Embedding-8B (Zhang et al., 2025) provides all embedding representations. The experiments ran on four NVIDIA A800 GPUs; the largest main runs took up to 12 hours. Tables 5, 10, and 6 use the fixed subsets defined below. Five judge runs reuse the same generated predictions and change only the matching-judge call. The LLM prompts used for the reported runs are reproduced in Appendix F.

A.2 Ablation Subsets

The RareArena subset contains at most 20 cases from each of 28 medical subfields, for 495 cases. The RareBench subset contains HMS (88 cases), LIRICAL (100), MME (40), PUMCH_ADM (75), and RAMEDIS (100), for 403 cases. The subsets were fixed before any ablation.

B Method Details

B.1 Case-Profile Retrieval

The case-profile branch first retrieves at most B_{prof} diseases by embedding similarity. Let the budgeted neighbors be $(d_1, \dots, d_B)$ with sorted scores $s_1 \geq \dots \geq s_B$, where $B \leq B_{\text{prof}}$. If $B \leq 2$, the branch returns all B neighbors. Otherwise, the adaptive prefix avoids adding a weak tail when the score curve has an elbow. We set a minimum prefix length:

$$t_{\min} = \lceil \sqrt{B} \rceil, \quad (16)$$

and compute candidate drops:

$$\Delta_i = s_i - s_{i+1}, \quad i = t_{\min}, \dots, B-1. \quad (17)$$

Let m_{Δ} be the median of these drops and let MAD_{Δ} be their median absolute deviation. If $i^* = \arg \max_i \Delta_i$ satisfies:

$$\Delta_{i^*} > m_{\Delta} + \text{MAD}_{\Delta}, \quad (18)$$

we keep the prefix $L_{\text{prof}} = (d_1, \dots, d_{i^*})$; otherwise we keep the full budgeted list $L_{\text{prof}} = (d_1, \dots, d_B)$. Thus the cutoff is applied only when the per-case similarity curve contains a drop that is unusually large relative to nearby drops; smooth curves retain the full retrieval budget.

B.2 Evidence State and Refinement Validation

Evidence sanitization. Structured HPO identifiers are canonicalized by extracting the first "HP: ..." pattern and preserving first occurrence order. Structured observed and negated HPO inputs receive weight 1, so they override duplicate free-text-derived entries for the same polarity. Free-text-derived HPO entries retain their normalizer similarity as reliability, clipped to $[0, 1]$.

If the same HPO term is present in both observed and negated evidence, evidence sanitization keeps only the polarity with the larger reliability. If the two reliabilities are equal, it keeps the observed entry and removes the negated entry.

Refinement validation. Refinements requested by the decider are suggestions rather than executable commands. ORBIT applies a refinement only if it appears in a programmatically generated shortlist, has canonical syntax, is not a duplicate, and can change the current candidate pool or evidence state. The accepted canonical forms are:

$$\begin{aligned} \text{BLOCK_DISEASE } \text{orpha}: \langle \text{code} \rangle, \\ \text{REQUIRE_MATCH } \text{hpo}: \langle \text{HP id} \rangle, \\ \text{IGNORE_OBSERVATION } \text{hpo}: \langle \text{HP id} \rangle. \end{aligned} \quad (19)$$

Scorer state. Let P^{base} and N^{base} be the sanitized observed and negated HPO sets. At iteration t , validation converts the accepted commands into three typed controls: a blocked-disease set $B_t \subseteq \mathcal{C}_t$, a required-term set $R_t \subseteq P^{\text{base}}$, and an ignored-term set $I_t \subseteq P^{\text{base}}$. The active evidence is:

$$P_t = P^{\text{base}} \setminus I_t, \quad N_t = N^{\text{base}}. \quad (20)$$

BLOCK_DISEASE adds a disease to B_t , REQUIRE_MATCH adds an observed term to R_t , and IGNORE_OBSERVATION adds an observed term to I_t . Evidence reliabilities are carried forward as q_h^+ and $q_h^- \in (0, 1]$. Together, these quantities form the scorer state:

$$\mathcal{S}_t = (\mathcal{C}_t, P_t, N_t, B_t, R_t, q^+, q^-). \quad (21)$$

Frequency label	lower	upper
Obligate (100%)	1.000	1.000
Very frequent (80% to 99%)	0.800	0.990
Frequent (30% to 79%)	0.300	0.790
Occasional (5% to 29%)	0.050	0.290
Very rare (1% to 4%)	0.010	0.040
Excluded (0%)	0.000	0.000

Table 8: Frequency-label intervals used to build disease-to-HPO parameters.

B.3 Probabilistic HPO Scoring

Notation. The scorer receives $\mathcal{S}_t$ from Eq. (21). The sanitized observed set P^{base} and initial retrieval-fused candidate set $\mathcal{C}_0$ remain fixed for the case. For each candidate disease $d \in \mathcal{C}_t$, the scorer converts catalog frequencies into disease-term probabilities, assigns a prior, and combines that prior with the positive and negative term likelihoods in Eq. (11). We write $\text{clip}(x, l, u) = \min\{u, \max\{l, x\}\}$.

Frequency parameters. Orphanet disease-to-HPO frequency labels are converted to the intervals in Table 8 before scoring.

For a non-degenerate frequency interval $[l, u]$ for an annotated disease-term pair (d, z) , we fit a Beta distribution by matching its 5th and 95th percentiles:

$$(\alpha_{dz}, \beta_{dz}) = \arg \min_{\alpha, \beta > 0} \left(Q_{.05}(\alpha, \beta) - l \right)^2 + \left(Q_{.95}(\alpha, \beta) - u \right)^2, \quad (22)$$

where Q_p is the p -quantile of $\text{Beta}(\alpha, \beta)$. For the degenerate boundary labels, set $\varepsilon_B = 10^{-6}$ and $\kappa_B = 200$, and define

$$m_{dz} = \begin{cases} \varepsilon_B, & [l, u] = [0, 0], \\ 1 - \varepsilon_B, & [l, u] = [1, 1], \end{cases} \quad (23)$$

$$(\alpha_{dz}, \beta_{dz}) = (\kappa_B m_{dz}, \kappa_B (1 - m_{dz})).$$

Thus both shape parameters remain positive and their sum is 200. For every frequency label, let $v_{dz} = \alpha_{dz} / (\alpha_{dz} + \beta_{dz})$ denote the base association frequency for annotated HPO term z under disease d .

Candidate prior. The ProbLog annotated disjunction contains the current candidate set $\mathcal{C}_t$ and an out-of-pool state. Let r_{unmap} be the number

of case-first LLM proposal names that remain unmapped to Orphanet. If the seed set $\mathcal{C}_0$ is non-empty, the prior probability of the out-of-pool state is:

$$p_{\text{other}} = \frac{1 + r_{\text{unmap}}}{|\mathcal{C}_0| + 1 + r_{\text{unmap}}}. \quad (24)$$

If $\mathcal{C}_0$ is empty, $p_{\text{other}} = 1$. For non-empty $\mathcal{C}_t$, each current candidate initially receives an equal share of $1 - p_{\text{other}}$. After blocking and guard adjustments, its effective prior is:

$$\pi_d = \frac{1 - p_{\text{other}}}{|\mathcal{C}_t|} \rho_d, \quad (25)$$

where $\rho_d \in [0, 1]$ is the candidate-specific prior multiplier. At each scorer call, $\rho_d = 0$ for a blocked disease $d \in B_t$ and $\rho_d = 1$ otherwise. The candidate-specific guard in Eq. (27) may then lower this value. Blocking and guard adjustments lower the affected candidate priors without increasing the priors of other diseases.

Candidate-specific guard. The guard reduces the prior of a candidate that has some observed support but is missing highly characteristic HPO terms. Let H_d be the annotated HPO terms for d , $f_{dh} = v_{dh}$, and $r_d(x)$ be the empirical rank percentile of x within $\{f_{dh} : h \in H_d\}$. Define:

$$\begin{aligned} m_d &= \text{median}_{h \in H_d} f_{dh}, \\ b_d &= \max(10^{-6}, m_d), \\ S_d &= \sum_{h \in P^{\text{base}} \cap H_d} f_{dh}, \\ c_{dh} &= \alpha_{dh} + \beta_{dh}, \\ u_{dh} &= \max\{r_d(f_{dh}) - r_d(m_d), 0\} \frac{c_{dh}}{c_{dh} + b_d}, \\ \chi_d &= \frac{S_d}{S_d + b_d} \sum_{h \in H_d \setminus P^{\text{base}}} u_{dh}. \end{aligned} \quad (26)$$

Here m_d is the disease’s median annotation frequency, b_d is that median clipped away from zero, S_d is the observed support score, and c_{dh} is the Beta concentration for term h . The term weight u_{dh} is positive only for an unobserved annotation above the disease’s median frequency percentile, and χ_d aggregates these missing annotations after weighting them by frequency percentile and Beta concentration. Larger values therefore indicate that more characteristic findings are absent from the case. The leading factor makes $\chi_d = 0$ when the case provides no observed support for d . Let $\mathcal{A}_t = \{d \in \mathcal{C}_t : \chi_d > 0\}$. An empty $\mathcal{A}_t$ leaves

all candidate priors unchanged. For nonempty $\mathcal{A}_t$, define $\tilde{\chi} = \text{median}\{\chi_d : d \in \mathcal{A}_t\}$. The guard then computes the downweighting factor γ_d and updates ρ_d as:

$$\begin{aligned} \rho_d &\leftarrow \rho_d \cdot \gamma_d, \\ \gamma_d &= \text{clip}\left(\frac{\tilde{\chi}}{\tilde{\chi} + \chi_d}, 10^{-4}, 1\right). \end{aligned} \quad (27)$$

Candidates with $\chi_d = 0$ are unchanged. The updated ρ_d enters Eq. (25).

Phenotype likelihood. For case HPO term h , let $A_d(h)$ be the set of annotated HPO terms a for disease d that imply h under ancestor closure, including h itself when it is directly annotated. For each $a \in A_d(h)$, the scorer uses the clipped association frequency $\theta_{da} = \text{clip}(v_{da}, 10^{-4}, 1 - 10^{-12})$.

To discount ancestor-descendant overlap within each evidence polarity, let $\mathcal{E}_t^+ = P_t$ and $\mathcal{E}_t^- = N_t$. For $h \in \mathcal{E}_t^s$, let k_h^s count the other terms $g \in \mathcal{E}_t^s$ for which either term is an HPO ancestor of the other, and define:

$$\omega_h^s = \frac{(1 + k_h^s)^{-1/2}}{\max_{g \in \mathcal{E}_t^s} (1 + k_g^s)^{-1/2}}. \quad (28)$$

For polarity $s \in \{+, -\}$, define:

$$\begin{aligned} T_s(p, \omega) &= \begin{cases} p^{\sqrt{\omega}}, & s = +, \\ 1 - (1 - p)^{\sqrt{\omega}}, & s = -, \end{cases} \quad (29) \\ \bar{\theta}_{da}^s &= T_s(\theta_{da}, \omega_a^s). \end{aligned}$$

where ω_a^s is the dependency weight in Eq. (28) and defaults to 1 when annotation a is not itself an active evidence term.

Let u_{bg} be the number of observed HPO terms with no direct catalog association in any seed-pool disease, $n_{cand} = |\mathcal{C}_0|$ the seed-pool size, and $n_{obs} = |P^{\text{base}}|$ the number of observed HPO terms. The per-case adaptive leak probability is:

$$\lambda = \begin{cases} 0, & u_{bg} = 0, \\ \frac{u_{bg}}{n_{cand} + n_{obs} + u_{bg}}, & u_{bg} > 0. \end{cases} \quad (30)$$

These three counts determine λ once for each case. The scorer uses this value in two places. It supplies the phenotype leak term for evidence left unexplained by the seed pool and the baseline probability of a spurious report. Let $\lambda_c = \text{clip}(\lambda, 10^{-12}, 1 - 10^{-12})$ and $\epsilon = 10^{-10}$. Using

the required-term control R_t in $\mathcal{S}_t$, we set:

$$\begin{aligned} \lambda_h^s &= \begin{cases} \epsilon, & h \in R_t, \\ T_s(\lambda_c, \omega_h^s), & h \notin R_t, \end{cases} \\ \eta_h &= \begin{cases} \epsilon, & h \in R_t, \\ \lambda_c, & h \notin R_t. \end{cases} \end{aligned} \quad (31)$$

Here λ_h^s is the polarity-adjusted leak probability, whereas η_h is the probability of a spurious report. Both are derived from the same per-case λ , except for terms in R_t , where they are set to ϵ . The effective phenotype probability for an evidence term of polarity s is:

$$y_{dh}^s = 1 - (1 - \lambda_h^s) \prod_{a \in A_d(h)} (1 - \bar{\theta}_{da}^s). \quad (32)$$

y_{dh}^s is the probability used by the likelihood for phenotype h under candidate d and polarity s . For $h \in R_t$, the background contribution is suppressed, so the term must be explained by catalog support.

Positive and negative reports are treated as noisy readouts of the latent phenotype. With reliabilities q_h^+ and q_h^- , the local likelihood factors are:

$$\begin{aligned} \ell_h^+(d) &= 1 - (1 - q_h^+ y_{dh}^+)(1 - \eta_h), \\ \ell_h^-(d) &= 1 - (1 - q_h^- (1 - y_{dh}^-))(1 - \eta_h). \end{aligned} \quad (33)$$

These factors are the terms used in the posterior score in Eq. (11).

C Hyperparameter Settings

For the LLM sampling parameters, we use the default values in each LLM’s config JSON file. Table 9 lists the fixed hyperparameters used in the main experiments unless otherwise stated for ablation. These parameters define candidate budgets, iterative refinement limits, and normalization thresholds.

`llm_topM`, `exact_retrieval_topk`, and `case_profile_topk` bound the number of candidates generated by the proposal LLM, direct HPO retrieval, and case-profile match, respectively. `problog_topk` controls how many ProbLog-ranked candidates are exposed as symbolic evidence, while `final_rerank_candidate_cap` limits the size of the fused pool shown to the final reranker. `max_iters` limits symbolic refinement. `extract_min_score` filters normalized HPO mentions from free text, `name_map_min_sim` filters disease-name grounding to Orphanet, and `free_text_min_observed_hpo` specifies the minimum accepted HPO count required to run ProbLog on free-text cases.

Hyperparameter	Value
llm_topM	20
exact_retrieval_topk	20
case_profile_topk	20
problog_topk	10
final_rerank_candidate_cap	30
max_iters	5
extract_min_score	0.8
name_map_min_sim	0.8
free_text_min_observed_hpo	10

Table 9: Hyperparameter settings used for the main ORBIT experiments.

D Judge Robustness

Table 10 compares DeepSeek-V3.2 (DS) with Gemini 3.5 Flash (GM) across five matching-judge runs on fixed predictions. The cross-system Spearman correlation between judges is 0.943 on RareArena and 1.000 on RareBench. Absolute scores shift, but the ordering and the leading system remain stable.

E RareArena Matching Details

Table 11 separates mutually exclusive score-1 hierarchical matches from score-2 equivalent-diagnosis matches on RareArena. Score 1 is not expected to rise with diagnostic quality because it is used only when score 2 is not met. Replacing a broad parent diagnosis with an equivalent disease-level answer moves a case from score 1 to score 2.

F LLM Prompt

This appendix presents the complete prompt text used in the experiments.

Phenotype extraction.

[SYSTEM]

You extract clinical phenotype mentions for HPO normalization. Use only the provided clinical text. Return STRICT JSON only in the format {"observed":[{"span": str}]}. Observed spans are explicit abnormalities, symptoms, signs, laboratory findings, imaging findings, developmental findings, or morphologic findings that are present. Do not extract absent, excluded, normal, denied, or otherwise negated findings. Do not include diagnoses, disease names, genes, treatments, procedures, medications, demographics, case identifiers, family history labels, or unstated inferences. Use exact short spans from the text when possible.

[USER]

CLINICAL TEXT: {clinical_text}

Open LLM proposal.

[SYSTEM]

You are a clinician generating a clinically precise differential diagnosis. Use only the provided input to produce medically plausible differential diagnoses. Rank candidates from most likely to least likely based on the evidence in the input. Do not use information not contained in the input. Return STRICT JSON only in the format {'candidates':[{'name': str, 'why': str}]}. The candidates list must be ordered from

highest to lowest likelihood. Return exactly {max_out} candidates.

[USER]

{proposal_input}

Refinement decider.

[SYSTEM]

You are the symbolic-refinement decider for whether another diagnostic iteration is worth the cost. Build your decision by reconciling the case evidence, the ProbLog/grounded evidence, and the case-first LLM differential. The case-first LLM differential is useful but not privileged. The grounded ProbLog candidates are useful but not automatically authoritative. If the symbolic state is coherent and no concrete patch or new grounded candidate would change it, choose STOP. If an unmapped case-first candidate is clinically important, preserve it conceptually in the analysis; it will be passed separately to the final reranker.

The decider input contains three evidence tracks:

1. CASE EVIDENCE: the primary clinical facts.
2. PROBLOG / GROUNDED EVIDENCE: compact catalog-grounded symbolic evidence from normalized HPO terms and the current ProbLog round. It supports stop/continue and patch decisions.
3. CASE-FIRST LLM DIFFERENTIAL: an independent open-ended differential diagnosis.

Only the non-obvious fields need special interpretation:

- CASE-FIRST CATALOG MAPPING: grounded means the disease passed the catalog name-mapping similarity threshold; unmapped means it did not.
- PROBLOG GROUNDED CANDIDATES posterior, when present: relative ProbLog score within the current grounded round only, not a calibrated real-world probability.
- POSTERIOR CATALOG CANDIDATES: compact catalog diseases retained by the current ProbLog posterior pass. This is not the final diagnosis, and unmapped open candidates are handled separately.
- support / conflict / missing: short clue lists for support, contradiction, and notable absent key features.
- AGREEMENT / DISAGREEMENT SUMMARY: compact comparison between the case-first differential and the current grounded round.
- PATCH CANDIDATES: ordered symbolic action shortlist focused on current disagreements. Earlier items are stronger suggestions, but they are still not precommitted patch decisions.
- OBSERVED HPO REFERENCE: broader lightweight list of observed HPO targets that may still be worth considering if a patch is needed.

Interpretation rules:

- The grounded ProbLog posterior, when present, only covers grounded catalog diseases. Do not treat an unmapped case-first disease as contradicted just because it has no grounded ProbLog posterior.
- Do not automatically preserve the case-first differential. Reconcile it against the case evidence and the grounded evidence.
- Do not automatically trust ProbLog either. Use the symbolic evidence to identify unresolved conflicts, missing evidence, and useful patches.
- When the case-first and grounded tracks disagree, use the agreement/disagreement summary and support/conflict/missing clues to decide whether another symbolic update is needed.
- Use patches mainly to resolve concrete disagreements between the case-first and grounded tracks, not just to chase generic curiosity.

Do not keep iterating just because the grounded ranker has a concentrated ranking if the case-level reasoning is already coherent. A CONTINUE decision is worthwhile only when you can name at least one concrete next-round action that will materially change the next grounded / ProbLog round: either one or more actionable PROBLOG patches, or one or more new candidates that can be grounded into the search pool. If your only new ideas are unmapped diseases that will not change the grounded search, preserve them in the eventual final differential instead of using them to justify CONTINUE. Do not repeat any diagnosis already present in the CASE-FIRST LLM DIFFERENTIAL. new_candidates must be truly new relative to that differential; mere rephrasing or re-listing is invalid. Unmapped names that are already in the CASE-FIRST LLM DIFFERENTIAL must not be repeated as new_candidates. If you cannot name a concrete patch or a concrete new grounded candidate, prefer STOP. Use AGREEMENT / DISAGREEMENT SUMMARY as the main bridge between the case-first and grounded tracks. Treat PATCH CANDIDATES as the only allowed source of PROBLOG patches; if none of those candidate actions is clearly

Method	RareArena		RareBench	
	DS	GM	DS	GM
DeepSeek-V3.2	0.4760 $\pm$ 0.0104	0.4549 $\pm$ 0.0061	0.5901 $\pm$ 0.0064	0.5767 $\pm$ 0.0060
Baichuan-M2	0.3370 $\pm$ 0.0063	0.3010 $\pm$ 0.0374	0.5434 $\pm$ 0.0094	0.5221 $\pm$ 0.0022
gpt-oss-20b	0.3345 $\pm$ 0.0076	0.3200 $\pm$ 0.0037	0.5129 $\pm$ 0.0099	0.5045 $\pm$ 0.0081
MedAgents	0.4424 $\pm$ 0.0102	0.4287 $\pm$ 0.0033	0.5602 $\pm$ 0.0048	0.5597 $\pm$ 0.0073
MDAgents	0.3661 $\pm$ 0.0069	0.3459 $\pm$ 0.0145	0.5124 $\pm$ 0.0090	0.5005 $\pm$ 0.0069
ORBIT	0.5406 $\pm$ 0.0023	0.5301 $\pm$ 0.0011	0.6670 $\pm$ 0.0103	0.6600 $\pm$ 0.0025

Table 10: Hit@10 mean $\pm$ standard deviation across five matching-judge runs on fixed predictions. DS is DeepSeek-V3.2 and GM is Gemini 3.5 Flash.

Method	Score 1	Score 2
DeepSeek-V3.2	0.0995	0.4167
Baichuan-M2	0.1071	0.3045
gpt-oss-20b	0.0828	0.3078
MedAgents	0.0687	0.3559
MDAgents	0.0840	0.3191
ORBIT	0.0662	0.4321
Qwen3-30B-A3B	0.0902	0.2951
MedAgents	0.0738	0.3336
MDAgents	0.0814	0.3135
ORBIT	0.0652	0.3915

Table 11: RareArena Hit@10 under mutually exclusive score-1 hierarchical and score-2 equivalent-diagnosis matches.

justified, do not output any patch. OBSERVED HPO REFERENCE is broader backup context for analysis, not permission to invent additional patch targets. Earlier shortlist items should be considered first, but whether to use a listed patch should still be decided by your own reasoning from the case and grounded evidence.

PROBLOG PATCH RULES (use sparingly, only if they would materially improve the next round):
1. REQUIRE_MATCH hpo:HP:xxxxxxx Use when an OBSERVED HPO is a near-obligate discriminator that should be mandatory for plausible diagnoses.
2. IGNORE_OBSERVATION hpo:HP:xxxxxxx Use when an observed finding is likely incidental, secondary, extraction noise, or too nonspecific and is distorting the grounded search.
3. BLOCK_DISEASE orpha:xxxx Use only when a grounded disease in the current pool is strongly contradicted by the case or is missing key obligate features. Never invent unsupported patch types. Never use leak/link override patch types.

Return STRICT JSON with keys:

```
{
  'decision': 'CONTINUE' | 'STOP',
  'analysis': str,
  'new_candidates': [{'name': str}],
  'patches': [str]
}
```

[USER]

CASE EVIDENCE (PRIMARY): {case_evidence}

PROBLOG / GROUNDED EVIDENCE: {grounded_evidence}

CASE-FIRST LLM DIFFERENTIAL: {formatted_llm_candidates}

CURRENT PROBLOG PATCHES: {current_patches}

DECISION TASK: Choose STOP or CONTINUE with concrete grounded-search actions. Return only the requested JSON keys.

Disease-card reranker.

[SOURCE_LEGEND]

Source labels: exact_retrieval means candidates matched directly by observed HPO terms; case_profile means candidates

retrieved by embedding similarity between the clinical case profile and disease profiles; llm_candidates or proposal_llm means case-first LLM differential names that were mapped to Orphanet; problog means candidates ranked by ProbLog posterior from normalized HPO evidence; catalog_pool means catalog candidates retained by the controller without a more specific source label. fusion_rank is the candidate's rank after catalog-pool ordering before this rerank step; branch_ranks lists branch-local ranks as branch:rank pairs, where a smaller rank is better inside that branch; use branch_ranks only as source metadata, not as authority for any single branch.

[SYSTEM]

You are a rare-disease diagnostic reranker. You must rank only the provided candidates for the given case. Catalog candidates are grounded to Orphanet; OPEN candidates are unmapped case-first LLM candidates. {SOURCE_LEGEND} Use the case information as primary evidence; use {secondary_metadata} as secondary evidence. Treat each source label, including problog, as one supporting branch without privileging one source over another. Do not invent diseases outside the candidate list. Return STRICT JSON only: {'panel10': [{'candidate_id': str, 'why': str}]}. Use candidate_id copied exactly from VALID CANDIDATE IDS, e.g. C1 or OPEN:1.

[USER]

CASE: {case_block}

CANDIDATES: {candidate_cards}

VALID CANDIDATE IDS (copy candidate_id exactly from this list): {id_menu}

Select and order exactly {topn} candidates if at least {topn} candidates are provided; if fewer than {topn} candidates are provided, return all provided candidates. Prefer candidates that explain the whole case from the CASE text, not just the normalized feature hints.

G Case study

Case analysis demonstrated the superiority of the ORBIT framework compared to the single LLM. For example, Figure 4 presents a case with a final diagnosis of congenital alpha-2-antiplasmin deficiency. In the single-LLM setting, the model mainly generated diagnoses related to platelet dysfunction and coagulation factor deficiencies in response to the patient's recurrent bleeding history. In contrast, within the ORBIT framework, HPO extraction captured the distinctive phenotype of prolonged bleeding after circumcision, which suggests dysfunction in the later stage of hemostasis, such as the fibrinolytic pathway. This key phenotype guided retrieval and reranking toward the correct final diagnosis. This case highlights the potential value of combining structured phenotype

Original Case

An 11-year-old Sudanese boy presented with frequent bleeding episodes, including bruising, nosebleeds, and gum bleeds. The most common type of hemorrhage was extensive bleeding after trauma, followed by subcutaneous bleeding. The patient had a history of delayed bleeding episodes since circumcision which took approximately one month to stop. Repeated episodes of nosebleeds were also reported. He did not receive blood or blood product transfusions. He reported subcutaneous bleeding following trauma while playing with friends. Hemorrhage into joints also occurred after trauma, but spontaneous joint bleeding was not reported. Pain and slight swelling of muscles sometimes persisted for months. Family history revealed that all family members were normal except his mother, who had a history of severe interval bleeding episodes (heavy menstrual bleeding). The patient had no known chronic illnesses, drug allergies, or developmental abnormalities. On examination, the patient was neither jaundiced nor cyanosed, and was not dysmorphic. There was no noticeable change when checking the ear, nose, and throat. The cardiopulmonary examination was normal. The patient's central nervous system examination showed good sensation and reflexes. He had no skin rash or organomegaly. Liver function tests and complete blood count were all normal. Viral screening for HIV, HBV, and HCV were negative.

Final diagnosis: Congenital alpha2-antiplasmin deficiency

Single LLM

von Willebrand disease type 1
Hemophilia A
Hemophilia B
Glanzmann thrombasthenia
Bernard-Soulier syndrome
MYH9-related disorder
Platelet function defect due to integrin deficiency / GPIIb/IIIa
Platelet function defect due to GPIb-vWF interaction disorder
Acquired severe platelet dysfunction
Combined factor VIII and vWF deficiency

Orbit: HPO extraction

HP:0001934 — Persistent bleeding after trauma
HP:0000978 — Bruising susceptibility
HP:0000421 — Epistaxis
HP:0000225 — Gingival bleeding
HP:0001933 — Subcutaneous hemorrhage
HP:0030137 — Prolonged bleeding following circumcision
HP:0005261 — Joint hemorrhage

HPO-retrieval candidate

1. Congenital factor V deficiency
2. Congenital factor XIII deficiency
3. Congenital alpha2-antiplasmin deficiency
4. Combined deficiency of factor V and factor VIII
5. Congenital factor VII deficiency
6. Familial hypofibrinogenemia
7. Severe hemophilia A
8. Acquired von Willebrand syndrome
9. Familial dysfibrinogenemia
10. Congenital factor X deficiency
11. Bleeding disorder in hemophilia A carriers
12. Attenuated Chédiak-Higashi syndrome
13. Congenital plasminogen activator inhibitor type 1 deficiency
14. Familial afibrinogenemia
15. Congenital factor XI deficiency
16. Gray platelet syndrome
17. Congenital factor II deficiency
18. Dengue fever
19. Glanzmann thrombasthenia
20. Moderate hemophilia A

LLM-candidates

1. Von Willebrand disease type 1
2. Von Willebrand disease type 2B
3. Von Willebrand disease type 3
4. Glanzmann thrombasthenia
5. Bernard-Soulier syndrome
6. Severe autosomal recessive macrothrombocytopenia / MYH9-related macrothrombocytopenia
7. Hemophilia A
8. Hemophilia B
9. Congenital factor XI deficiency
10. Congenital factor VII deficiency
11. Congenital factor X deficiency
12. Congenital factor V deficiency
13. Congenital factor II deficiency
14. Acquired factor XIII deficiency
15. Gray platelet syndrome
16. Familial dysfibrinogenemia
17. Familial hypofibrinogenemia
18. Acquired von Willebrand syndrome
19. Bleeding diathesis due to glycoprotein VI deficiency
20. Autoimmune thrombocytopenia

Case-retrieval candidate

1. Congenital factor V deficiency
 2. Congenital alpha2-antiplasmin deficiency
 3. Moderate hemophilia A
 4. Severe hemophilia A
 5. Hemophilia B
 6. Hemophilia A
 7. Bleeding disorder in hemophilia A carriers
 8. Glanzmann thrombasthenia
- Final answer**
- Moderate hemophilia A
Hemophilia A
Hemophilia B
Severe hemophilia A
Congenital alpha2-antiplasmin deficiency
Congenital factor V deficiency
Congenital factor X deficiency
Congenital factor XIII deficiency
Congenital factor XI deficiency
Congenital factor VII deficiency

Physician Analysis:

For patients with recurrent bleeding, several hemostatic components should be considered, including platelet number or function, coagulation factor deficiencies, and, less commonly, disorders of fibrinolysis. In the current case, the bleeding phenotype is mixed, with both mucocutaneous or subcutaneous bleeding and deeper joint bleeding after trauma. Therefore, it is reasonable that the model initially considered both platelet function disorders and coagulation factor deficiencies. However, the most informative clue is the markedly prolonged bleeding after circumcision, which suggests a delayed bleeding process related to impaired clot stability or excessive fibrinolysis. This explains why alpha-2-antiplasmin deficiency, a representative disorder of impaired fibrinolytic inhibition, was retrieved as a relevant candidate.

Figure 4

extraction, retrieval, and reranking to reduce noise from free-text narratives and focus on diagnostically informative but sparsely described clinical features.

H Consent and Use of Data

The physician reviewers were recruited by direct invitation based on their clinical expertise. No crowdsourcing platform, student participant pool, or public recruitment process was used. The reviewers did not receive financial compensation for this review. The clinical cases used in this study were obtained from a publicly available, open-source dataset and had been deidentified before use. The study did not involve direct interaction with the patients or collection of new patient data. No protected health information or personally identifiable information was included in the materials reviewed by the physician.

Because the cases involved were publicly available and deidentified, the authors did not obtain additional consent for the cases. The cases were used only for research evaluation of the LLM-related system. The physician reviewers were informed that the cases were publicly available and deidentified, and that the review was conducted solely for research purposes rather than for patient care. The use of the cases was consistent with the publicly available nature of the source dataset and its deidentified status. No attempt was made to reidentify any individual.

I Potential Risks and Risk Mitigation

The potential risks of this work were limited. All clinical cases were obtained from publicly available, deidentified datasets, and all models or LLM-related systems used in the study were publicly available. The study did not involve private patient records, protected health information, direct patient contact, or clinical decision-making.

The main potential risks were theoretical residual reidentification risk, possible misinterpretation of LLM-related outputs as clinical guidance, and possible amplification of errors or biases present in the public dataset or model outputs. To reduce these risks, the study did not attempt to reidentify any individual, did not use the outputs for patient care, and did not redistribute the original dataset or model materials. The physician review was conducted solely for research evaluation purposes.