HPOQuest: A Rare-Disease Diagnostic Agent Using Active Phenotype Acquisition

Kamilia Zaripova[1,2,3,4], Nassir Navab[1,3], Azade Farshad[1,3,5†], and Annalisa Marsico[2†]

[1] Technical University of Munich, Munich, Germany
[2] Computational Health Center, Helmholtz Center Munich, Oberschleißheim, Germany
[3] Munich Center for Machine Learning (MCML), Munich, Germany
[4] Munich Data Science Institute (MDSI), Munich, Germany
[5] Aalto University, Espoo, Finland
kamilia.zaripova@tum.de

Abstract. More than 300 million people worldwide are affected by one of over 7,000 known rare diseases, yet diagnosis remains difficult because patients initially present with incomplete and heterogeneous phenotypes. We present **HPOQuest**, a training-free framework for sequential phenotype acquisition in rare-disease diagnosis. Starting from a small set of observed patient phenotypes, HPOQuest maintains a probabilistic disease ranking and iteratively selects informative follow-up questions to support clinicians during patient assessment. Confirmed phenotypes update the disease ranking, while all responses update the candidate question set. Across four benchmark cohorts, HPOQuest substantially improves diagnosis from sparse initial phenotypes, with gains of up to 30% points at Recall@1 and 45% points at Recall@5. These results demonstrate that sequential phenotype acquisition can substantially improve rare-disease diagnosis from limited initial clinical evidence.

Keywords: Rare disease diagnosis, Sequential phenotype acquisition, Human Phenotype Ontology, Active information acquisition

1 Introduction

Although individually rare, each affecting fewer than 1 in 2,000 people in Europe, rare diseases collectively affect more than 300 million people worldwide and remain difficult to diagnose. Clinicians typically begin with a few signs and symptoms and refine the differential through targeted questions, examinations, and tests. Patient findings are commonly represented using the Human Phenotype Ontology (HPO), a hierarchy of more than 20,000 phenotypic abnormalities. However, extensive overlap among diseases and variation among patients with the same disease leave many diagnoses plausible from only a few initial terms.

† Shared last authorship

These terms also vary in diagnostic value: specific findings may identify the relevant disease family, whereas common, nonspecific, or incidental findings may provide little guidance.

Current rare-disease systems primarily diagnose from observed phenotype profiles or augment diagnosis with retrieval and LLM reasoning, but do not explicitly address sequential, response-dependent phenotype acquisition. Existing rare-disease methods primarily model disease–phenotype associations or generate diagnoses assuming fully supplied phenotype profiles [3,24]. RareBench [4] benchmarks phenotype extraction and rare-disease diagnosis, while systems such as DeepRare [24] and RDguru [23] combine LLMs with phenotype processing, retrieval, and diagnostic tools. DeepRare diagnoses from a supplied phenotype profile, whereas RDguru performs interactive phenotype acquisition but starts from 50% of the recorded phenotypes and relies on a multi-component GPT-4-based agent. PhenoDP [21] recommends one additional phenotype from partially observed profiles without modeling sequential patient-dependent questioning.

Clinical questioning has been studied by MediQ and MedKGI through free-form question asking and information-guided test selection [12,20]. MedClarify [22] similarly selects free-form questions using expected information gain. LACDM [2] learns cost-efficient test selection using reinforcement learning within four abdominal diseases, while other systems learn diagnostic trajectories or employ specialized diagnostic agents [10,17]. Earlier active feature acquisition methods rely on model attribution [19] or information-theoretic selection [9], whereas ASIG [8] learns an LLM-based acquisition policy on 20 Questions and transfers it to MediQ. Together, these methods address related information-acquisition problems but not sequential, response-dependent phenotype selection over the HPO hierarchy across thousands of rare diseases.

Selecting the next phenotype remains challenging because the HPO contains more than 20,000 hierarchically related terms, diseases share overlapping phenotype profiles, and each response changes both the candidate diseases and the valid follow-up questions. HPOQuest addresses this setting by starting from the same three seed phenotypes for every patient, compared with median complete-profile sizes of 9–17.5 HPO terms. It iteratively selects a limited number of valid HPO questions, and updates both the candidate diseases and the available follow-up questions after each response. Its training-free acquisition policy prioritizes questions using the current candidate diseases, disease–phenotype association frequencies, and phenotype specificity.

HPOQuest uses a locally deployable, in-clinic, open-weight model and compares all question selection methods using the same inputs, budget, responses, retrieval, and diagnostic components. We contribute (i) an ontology-constrained framework for phenotype acquisition from retrospective HPO records, where three-way responses expand the candidate question set toward confirmed descendants or prune absent subtrees; (ii) a training-free, candidate-anchored policy for hierarchical phenotype acquisition; and (iii) a controlled benchmark demonstrating when active phenotype acquisition improves rare-disease diagnosis. When sparse no-acquisition diagnosis is already informative, HPOQuest remains com-

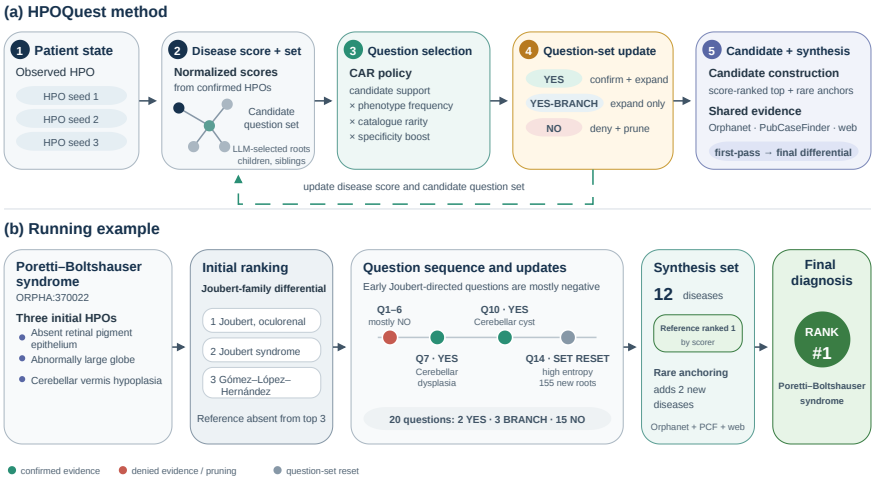

Fig. 1. HPOQuest. (a) Starting from three seed HPO phenotypes, HPOQuest iteratively updates disease scores and the ontology-constrained candidate question set. The candidate-anchored rarity (CAR) policy selects follow-up questions, responses update both the disease ranking and candidate set, and the acquired phenotypes are combined with shared biomedical evidence to produce the final ranked differential diagnosis. (b) Representative diagnostic trajectory for a patient with Poretti–Boltshauser syndrome, showing how sequential phenotype acquisition eliminates competing diagnoses and promotes the reference disease to rank 1.

petitive with the strongest baseline across all evaluated cutoffs. When both sparse baselines fail, as on the MME cohort, HPOQuest improves recall by 30.0–45.0 percentage points. HPOQuest therefore complements accessible LLM and retrieved knowledge while providing its largest gains when that knowledge is insufficient.

2 Method

The Human Phenotype Ontology (HPO) is a directed acyclic graph whose nodes represent clinical abnormalities and whose edges connect general phenotypes to more specific ones. HPOQuest uses these terms as candidate questions. Starting from a small set of observed phenotypes, it repeatedly ranks diseases, selects valid follow-up questions, and updates both the disease ranking and candidate question set after each response. The acquired phenotypes are then combined with retrieved biomedical evidence to produce a ranked differential diagnosis.

Patient state and disease posterior. HPOQuest begins with only K observed seed phenotypes and iteratively acquires additional evidence. Confirmed phenotypes are used to rank diseases, and the ranking is updated after every new confirmation to guide question selection. Let $\mathcal{H}$ denote the set of HPO terms and

$\mathcal{D}$ the disease catalogue. Each disease $d \in \mathcal{D}$ is associated with an HPO profile profile(d) and phenotype frequencies $P(h \mid d)$. For patient p, $\mathcal{H}^+(p)$ denotes the complete recorded positive phenotype set, with $\mathcal{S} \subseteq \mathcal{H}^+(p)$ the initially observed seed phenotypes. HPOQuest may ask at most B additional phenotype questions. During evaluation, responses are generated deterministically from the recorded patient profile:

$$a(h) = \begin{cases} \text{YES}, & h \in \mathcal{H}^+(p), \\ \text{YES-BRANCH}, & h \notin \mathcal{H}^+(p) \text{ and } \mathrm{Desc}(h) \cap \mathcal{H}^+(p) \neq \emptyset, \\ \text{NO}, & \text{otherwise.} \end{cases} \tag{1}$$

Because the response simulator is defined only with respect to the recorded positive phenotype set $\mathcal{H}^+(p)$, it adopts a closed-world assumption: terms absent from $\mathcal{H}^+(p)$, with no recorded descendant, are treated as negative. Here, Desc(h) contains the more specific HPO terms below h. A YES confirms h; YES-BRANCH indicates a recorded descendant; and NO denies h. We maintain confirmed and denied sets $\mathcal{C}$ and $\mathcal{N}$, initialized as $\mathcal{C} = \mathcal{S}$ and $\mathcal{N} = \emptyset$; a YES-BRANCH term is added to neither set.

Confirmed phenotypes contribute weighted evidence for associated diseases. The weight w_h downweights generic seed phenotypes, while phenotypes confirmed during questioning always receive unit weight:

We first measure how widely phenotype h occurs across disease profiles:

$$\hat{p}(h) = \frac{|\{d \in \mathcal{D} : h \in \mathrm{profile}(d)\}|}{|\mathcal{D}|}, \qquad w_h = \begin{cases} \gamma, & h \in \mathcal{S} \text{ and } \hat{p}(h) > \theta_s, \\ 1, & \text{otherwise.} \end{cases} \tag{2}$$

Here, $\hat{p}(h)$ is the fraction of diseases associated with h, θ_s identifies generic seed phenotypes, and $\gamma \in (0, 1)$ is their downweighting factor. If all seeds exceed θ_s, the least common seed retains unit weight. Phenotypes confirmed during questioning always receive unit weight. The phenotype frequencies are combined into a weighted log-compatibility score $\ell(d \mid \mathcal{C})$, where a higher value indicates that disease is more consistent with the currently confirmed phenotypes:

$$\ell(d \mid \mathcal{C}) = \sum_{h \in \mathcal{C}} w_h \log P(h \mid d), \qquad \pi(d \mid \mathcal{C}) = \frac{\exp(\ell(d \mid \mathcal{C}))}{\sum_{d' \in \mathcal{D}} \exp(\ell(d' \mid \mathcal{C}))}. \tag{3}$$

The softmax normalization converts these compatibility scores into a distribution $\pi(d \mid \mathcal{C})$ over the disease catalogue, which is used to rank diseases and guide question selection. Only confirmed phenotypes modify this score; denied phenotypes are used for pruning questions and during final diagnostic synthesis.

Ontology-constrained question selection. At step t, the candidate question set $\mathcal{F}_t \subset \mathcal{H}$ contains the HPO terms that may be queried next. It is initialized with the children and siblings of the seed terms. To cover relevant regions beyond these local branches, one LLM call selects additional top-level HPO categories

from a fixed list, excluding categories that already contain a seed. The LLM cannot generate new HPO terms or select subsequent questions.

Before questioning, we issue one web search for each of the K seed phenotypes. Question selection itself uses only the HPO structure and the disease–phenotype associations and frequencies. Let $\mathcal{T}_M$ be the M highest-ranked diseases under $\pi(d \mid \mathcal{C})$. We quantify the catalogue rarity of phenotype h by

$$r(h) = -\log \max\{\hat{p}(h), \phi\}, \tag{4}$$

and identify associations that are frequent within a disease but uncommon across the catalogue:

$$z(h,d) = \begin{cases} 1, & P(h \mid d) \geq \theta_f \text{ and } \hat{p}(h) \leq \theta_r, \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

The acquisition score combines the current disease ranking, within-disease phenotype frequency, catalogue rarity, and a specificity boost for phenotypes that are common within a disease but rare overall. The hyperparameters $\lambda \geq 0$ and $\beta \geq 0$ control the influence of the disease ranking and specificity boost. Our candidate-anchored rarity (CAR) policy assigns each eligible question $h \in \mathcal{F}_t$ the score

$$s(h) = \sum_{d \in \mathcal{T}_M} \underbrace{(1 + \lambda \pi(d \mid \mathcal{C}))}_{\text{current disease rank}} \underbrace{P(h \mid d)}_{\text{within-disease frequency}} \underbrace{r(h)}_{\text{catalogue rarity}} \underbrace{(1 + \beta z(h,d))}_{\text{specificity boost}}. \tag{6}$$

Higher scores identify phenotypes that are both characteristic of the current candidate diseases and discriminative across the disease catalogue. Questions with $s(h) \geq \tau$ are selected in descending score order, up to $\min\{b, B - q\}$, where b is the per-round limit and q the number of questions already issued.

Question-set update and stopping. A YES or YES-BRANCH response adds the queried term's children to $\mathcal{F}_t$. A NO response removes the queried term and all its descendants. Queried and removed terms cannot be selected again.

Let $\mathcal{T}_L$ contain the L highest-ranked diseases and let $\bar{\pi}(d) = \pi(d) / \sum_{d' \in \mathcal{T}_L} \pi(d')$. We measure uncertainty among these diseases as

$$H_L = -\sum_{d \in \mathcal{T}_L} \bar{\pi}(d) \log \bar{\pi}(d). \tag{7}$$

Questioning stops when the budget is exhausted or when $H_L < \eta_{\text{stop}}$ after at least $C_{\min}$ phenotypes have been confirmed. The question set may be reset once to branches not covered by the seeds if the top three diseases remain unchanged for three rounds without a new confirmation, or if $H_L > \eta_{\text{reset}}$ after Q_{reset} questions.

Candidate construction and diagnostic synthesis. Because the ranking induced by $\pi(d \mid \mathcal{C})$ may under-rank diseases supported by few rare phenotypes,

Table 1. Overall diagnosis results on the four RareBench cohorts (%).

	MME ($n = 40$)			HMS ($n = 88$)			LIRICAL ($n = 370$)			RAMEDIS ($n = 624$)		
Method	R@1	R@3	R@5	R@1	R@3	R@5	R@1	R@3	R@5	R@1	R@3	R@5
Complete phenotype profile: all recorded HPO terms												
PhenoBrain[15]	25.0	45.0	80.0	22.4	32.9	42.4	27.1	44.7	57.7	20.7	38.3	55.8
PubCaseFinder[6]	40.0	57.5	57.5	7.3	13.4	15.9	37.8	47.0	48.9	25.8	32.6	40.0
GPT-4o[11]	2.5	5.0	10.0	47.7	60.2	65.9	31.1	42.2	48.9	35.5	52.2	61.1
DeepSeek R1[7]	35.0	60.0	65.0	38.6	53.4	60.2	36.8	45.1	47.6	33.2	47.6	53.2
DeepRare (GPT-4o)	42.5	57.5	57.7	50.0	62.5	64.8	39.5	49.2	52.4	36.9	46.4	55.4
DeepRare (Qwen-32B)	17.5	22.5	35.0	25.0	38.6	47.7	39.5	47.6	51.9	45.0	61.2	65.4
Sparse phenotype profile: three initial HPO terms												
DeepRare (Qwen-32B)	5.0	5.0	5.0	12.5	18.2	22.7	**28.6**	35.1	39.2	**33.2**	45.0	48.9
Qwen-32B + web search	2.5	5.0	5.0	**18.2**	20.5	**29.5**	27.0	34.9	39.5	29.8	44.7	**52.6**
HPOQuest (Qwen-32B)	**35.0**	**42.5**	**50.0**	15.9	**23.9**	25.0	**28.6**	**40.3**	**45.1**	27.6	**45.4**	49.0

we compute

$$\text{anchor}(d) = \sum_{\substack{h \in \mathcal{C} \\ \hat{p}(h) \le \theta_a}} r(h) P(h \mid d). \tag{8}$$

Disease–phenotype associations and frequencies are obtained from Orphanet [16] and HPOA [14]; OMIM-only diseases [1] without reported phenotype frequencies are assigned a fixed probability. The synthesis set contains the top T diseases under π, up to A previously unseen diseases ranked by anchor(d), and the remaining π-ranked diseases, retaining the first R unique entries.

After candidate construction, an LLM-generated web query and a PubCaseFinder [6] query are issued from the confirmed phenotype set. Shared retrieval evidence and each candidate's Orphanet phenotype profile are then provided to two LLM passes, which produce a draft of size N_{draft} and a final ranking of size N_{final}, which may include diseases outside the synthesis set.

3 Experimental Setup

Datasets, baselines, and policies. We evaluate on the four RareBench [4] cohorts: MME ($n = 40$), HMS ($n = 88$), LIRICAL ($n = 370$), and RAMEDIS ($n = 624$). All sparse-input methods receive the same three seed phenotypes. For patients with more than three recorded phenotypes, we select the two phenotypes occurring in the fewest Orphanet disease profiles and sample one additional phenotype uniformly from the remaining terms using a deterministic patient-specific random seed; patients with at most three recorded phenotypes use all available terms. DeepRare [24] receives either these seeds or the complete phenotype profile. Qwen-32B + web receives only the seeds and web access, without phenotype acquisition or candidate retrieval. All controlled systems use the same diagnoser and judge, with DeepRare's private similar-case retrieval disabled. DeepRare with the complete phenotype profile and published full-profile results serve only as references.

Table 2. Acquisition-policy ablation on the four RareBench cohorts (%).

	MME ($n = 40$)			HMS ($n = 88$)			LIRICAL ($n = 370$)			RAMEDIS ($n = 624$)		
Configuration	R@1	R@3	R@5	R@1	R@3	R@5	R@1	R@3	R@5	R@1	R@3	R@5
No acquisition, Qwen-32B, 3 HPOs												
DeepRare	5.0	5.0	5.0	12.5	18.2	22.7	28.6	35.1	39.2	33.2	45.0	48.9
LLM + web search	2.5	5.0	5.0	18.2	20.5	29.5	27.0	34.9	39.5	29.8	44.7	52.6
HPOQuest acquisition policies												
LLM-guided selection	27.5	40.0	42.5	**20.5**	**28.4**	**30.7**	30.0	39.5	43.8	26.4	43.4	47.9
EIG	22.5	37.5	37.5	18.2	**28.4**	**30.7**	**31.4**	39.5	43.0	27.9	43.6	47.4
CAR	**35.0**	42.5	**50.0**	15.9	23.9	25.0	28.6	**40.3**	**45.1**	27.6	**45.4**	49.0
RRF	**35.0**	**47.5**	47.5	**20.5**	26.1	28.4	26.5	37.8	43.0	**29.0**	42.8	**49.7**

We compare CAR with three acquisition policies under the same experimental setting. LLM-guided selection inspired by MediQ [12] prompts an expert LLM to select from the current HPO candidate set or abstain. One-step expected information gain (EIG) [13] ranks candidate terms by their expected entropy reduction over the current top-M diseases using $P(h \mid d)$; only the ranking policy changes, while responses, including YES-BRANCH, follow Section 2. CAR–EIG combines the CAR and EIG rankings using reciprocal rank fusion (RRF) [5] with $\kappa = 60$.

Implementation details. We use $K = 3$ seed phenotypes, question budget $B = 20$, per-round limit $b = 8$, and $M = 50$ diseases for question scoring. Stopping uses $(L, \eta_{\text{stop}}, C_{\min}, \eta_{\text{reset}}, Q_{\text{reset}}) = (10, 0.8, 4, 1.2, 10)$. CAR uses $(\theta_s, \gamma, \lambda, \beta, \phi, \tau, \theta_f, \theta_r, \theta_a) = (0.02, 0.7, 4, 2, 0.002, 0.02, 0.55, 0.02, 0.01)$. Candidate construction uses $(T, A, R) = (8, 4, 12)$. Missing HPOA frequencies are assigned $P(h \mid d) = 0.5$. Anchor retrieval covers 4,283 Orphanet and 9,120 HPOA diseases. The synthesis passes use $(N_{\text{draft}}, N_{\text{final}}) = (12, 15)$. All controlled systems use 4-bit Qwen2.5-32B-Instruct on one GPU, SerpAPI for web search, and GPT-4o-mini with DeepRare's judging prompt. Runtime per patient is 370 s (CAR), 384 s (CAR–EIG), and 496 s/611 s (DeepRare with three/all phenotypes). We report Recall@k ($k \in \{1, 3, 5\}$), accepting predictions verified by either the LLM judge or a MONDO [18] synonym.

4 Results and Discussion

Overall performance. Active phenotype acquisition provides its clearest benefit on MME (Table 1), where HPOQuest with CAR improves Recall@1/3/5 from 5.0/5.0/5.0% to 35.0/42.5/50.0%, exceeding the Qwen-32B full-profile DeepRare reference despite starting from only three seed phenotypes. This indicates that acquiring a small number of informative phenotypes can be more effective than relying on the complete recorded profile when the initial evidence is insufficient for diagnosis. On LIRICAL, CAR improves Recall@3 and Recall@5 while matching DeepRare-3 at Recall@1. Gains are smaller on HMS and RAMEDIS, where the strong sparse-input baselines, particularly Qwen-32B + web, suggest that the language model can already infer a plausible disease region from the initial

Table 3. Results stratified by seed recoverability (%). R denotes seed-recoverable cases and NR non-recoverable cases. All configurations use Qwen-32B.

			DeepRare-All			DeepRare-3			Qwen-32B + web			HPOQuest (CAR)		
Cohort	Group	n	R@1	R@3	R@5	R@1	R@3	R@5	R@1	R@3	R@5	R@1	R@3	R@5
MME	R	9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	77.8	77.8	88.9
	NR	31	22.6	29.0	45.2	6.5	6.5	6.5	3.2	6.5	6.5	22.6	32.3	38.7
HMS	R	7	28.6	42.9	57.1	14.3	28.6	28.6	0.0	14.3	28.6	0.0	57.1	57.1
	NR	81	24.7	38.3	46.9	12.3	17.3	22.2	19.8	21.0	29.6	17.3	21.0	22.2
LIRICAL	R	74	54.1	62.2	68.9	39.2	47.3	56.8	43.2	58.1	66.2	47.3	73.0	81.1
	NR	296	35.8	43.9	47.6	26.0	32.1	34.8	23.0	29.1	32.8	24.0	32.1	36.1
RAMEDIS	R	114	52.6	76.3	80.7	46.5	78.1	78.1	44.7	68.4	86.0	40.4	87.7	91.2
	NR	510	43.3	57.8	62.0	30.2	37.6	42.4	26.5	39.4	45.1	24.7	35.9	39.6

phenotypes. In these cohorts, active acquisition primarily refines the differential diagnosis, yielding more consistent improvements at Recall@3 and Recall@5 than at Recall@1.

Acquisition policies. No single acquisition policy dominates across cohorts (Table 2). CAR performs best at deeper cutoffs on MME and LIRICAL, RRF achieves the highest Recall@3 on MME and Recall@1/5 on RAMEDIS, while EIG and LLM-guided selection perform best on HMS. These trends reflect the complementary objectives of the policies: CAR prioritizes phenotype specificity within the current candidate set, EIG explicitly reduces posterior uncertainty, and LLM-guided selection leverages the language model's prior biomedical knowledge. The results therefore suggest that the optimal acquisition strategy depends on the diagnostic setting rather than one policy being uniformly superior.

Seed recoverability. Table 3 stratifies patients according to seed recoverability, defined as whether the reference disease appears among the top-15 Orphanet candidates from the three seed phenotypes alone. Acquisition is most effective for seed-recoverable cases, raising Recall@1 from 0.0% to 77.8% on MME and from 39.2% to 47.3% on LIRICAL, with larger gains at deeper cutoffs. For non-recoverable cases, improvements remain limited across cohorts, suggesting that active questioning is most effective when the initial seed phenotypes already provide sufficient signal to identify a plausible diagnostic region. All reported 0.0% entries are observed results, indicating that no patient in the corresponding subgroup was ranked. Taken together, the results reveal two operating regimes. When the initial phenotypes provide only a weak signal for the language model, as on MME, active phenotype acquisition substantially improves diagnostic accuracy. In contrast, when the language model can already infer a plausible disease region from the seed phenotypes, as suggested by the strong Qwen-32B + web baselines on HMS, LIRICAL, and RAMEDIS, acquisition primarily refines the ordering of the differential diagnosis rather than changing the top-ranked prediction.

5 Conclusion

We introduced HPOQuest, a training-free framework for active HPO acquisition that formulates rare-disease diagnosis as sequential selection of ontology-constrained phenotype questions. Across four RareBench cohorts, HPOQuest substantially improves diagnosis when sparse initial phenotypes provide limited diagnostic evidence, while remaining competitive when the initial phenotype signal is already informative. Our analyses further show that the effectiveness of active acquisition depends on both the informativeness of the initial phenotypes and the strength of the language model's prior knowledge, suggesting that structured questioning primarily complements existing LLM capabilities by refining the differential diagnosis. Future work should investigate robustness across different initial phenotype settings and language models.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Amberger, J.S., Bocchini, C.A., Scott, A.F., Hamosh, A.: Omim. org: leveraging knowledge across phenotype–gene relationships. Nucleic acids research **47**(D1), D1038–D1043 (2019)
2. Bani-Harouni, D., Pellegrini, C., Özsoy, E., Navab, N., Keicher, M.: Language agents for hypothesis-driven clinical decision making with reinforcement learning. In: The Fourteenth International Conference on Learning Representations (2025)
3. Chen, S., Nguyen, Q.M., Hu, Y., Liu, C., Weng, C., Wang, K.: Phenoss: Phenotype semantic similarity-based approach for rare disease prediction and patient clustering. medRxiv (2026)
4. Chen, X., Mao, X., Guo, Q., Wang, L., Zhang, S., Chen, T.: 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)
5. Cormack, G.V., Clarke, C.L., Buettcher, S.: 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. pp. 758–759 (2009)
6. Fujiwara, T., Yamamoto, Y., Kim, J.D., Buske, O., Takagi, T.: Pubcasefinder: A case-report-based, phenotype-driven differential-diagnosis system for rare diseases. The American Journal of Human Genetics **103**(3), 389–399 (2018)
7. Guo, D., Yang, D., Zhang, H., Song, J., Wang, P., Zhu, Q., Xu, R., Zhang, R., Ma, S., Bi, X., et al.: Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948 (2025)
8. Hartmann, J., Harvey, J., Navott, J., Wang, E.Y., Melo, L.C., Cipcigan, F., Zhang, C., Abate, A.: Amortising bayesian experimental design for sequential information gathering in llms. arXiv preprint arXiv:2607.03426 (2026)
9. He, W., Mao, X., Ma, C., Huang, Y., Hernàndez-Lobato, J.M., Chen, T.: Bsoda: a bipartite scalable framework for online disease diagnosis. In: Proceedings of the ACM Web Conference 2022. pp. 2511–2521 (2022)

10. Hsu, H.L., Wang, Z., Zhang, D., Chen, N.C., Wang, J., Ding, J.E., Hsu, C.H., Wang, G., Liu, F., Hung, F.M., et al.: Medaction: Towards active multi-turn clinical diagnostic llms. arXiv preprint arXiv:2605.07305 (2026)
11. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024)
12. Li, S.S., Balachandran, V., Feng, S., Ilgen, J.S., Pierson, E., Koh, P.W., Tsvetkov, Y.: Mediq: Question-asking llms and a benchmark for reliable interactive clinical reasoning. Advances in Neural Information Processing Systems **37**, 28858–28888 (2024)
13. Lindley, D.V.: On a measure of the information provided by an experiment. The Annals of Mathematical Statistics **27**(4), 986–1005 (1956)
14. Ma, G., NM, B., et al.: The human phenotype ontology in 2024: phenotypes around the world. Nucleic Acids Res [Internet] **52**, D1 (2024)
15. Mao, X., Huang, Y., Jin, Y., Wang, L., Chen, X., Liu, H., Yang, X., Xu, H., Luan, X., Xiao, Y., et al.: A phenotype-based ai pipeline outperforms human experts in differentially diagnosing rare diseases using ehrs. NPJ Digital Medicine **8**(1), 68 (2025)
16. Rath, A., Olry, A., Dhombres, F., Brandt, M.M., Urbero, B., Ayme, S.: 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)
17. Sanghvi, A., Akash, N., Imam, R., Sharma, A., Jain, M.: Medxagent: Multi-agent consultation for interactive medical diagnosis. arXiv preprint arXiv:2606.03416 (2026)
18. Vasilevsky, N., Essaid, S., Matentzoglu, N., Harris, N.L., Haendel, M., Robinson, P., Mungall, C.J.: Mondo disease ontology: harmonizing disease concepts across the world. In: CEUR Workshop Proceedings, CEUR-WS. vol. 2807, pp. 1–2 (2020)
19. Vivar, G., Mullakaeva, K., Zwergal, A., Navab, N., Ahmadi, S.A.: Peri-diagnostic decision support through cost-efficient feature acquisition at test-time. In: International conference on medical image computing and computer-assisted intervention. pp. 572–581. Springer (2020)
20. Wang, Q., Sheng, R., Li, Y., Qu, H., Sun, Y., Zhu, M.: Medkgi: Iterative differential diagnosis with medical knowledge graphs and information-guided inquiring. arXiv preprint arXiv:2512.24181 (2025)
21. Wen, B., Shi, S., Long, Y., Dang, Y., Tian, W.: Phenodp: leveraging deep learning for phenotype-based case reporting, disease ranking, and symptom recommendation. Genome Medicine **17**(1), 67 (2025)
22. Wong, H.M., Heesen, P., Janetzky, P., Bendszus, M., Feuerriegel, S.: Medclarify: An information-seeking ai agent for medical diagnosis with case-specific follow-up questions. arXiv preprint arXiv:2602.17308 (2026)
23. Yang, J., Shu, L., Duan, H., Li, H.: Rdguru: a conversational intelligent agent for rare diseases. IEEE Journal of Biomedical and Health Informatics **29**(9), 6366–6378 (2024)
24. Zhao, W., Wu, C., Fan, Y., Qiu, P., Zhang, X., Sun, Y., Zhou, X., Zhang, S., Peng, Y., Wang, Y., et al.: An agentic system for rare disease diagnosis with traceable reasoning. Nature **651**(8106), 775–784 (2026)