HyperRAG: Combining RAG and Hyperbolic Embeddings for Phenotypes Linking from Text

Anonymous ACL submission

Abstract

Extracting knowledge from unstructured data is a critical task for advancing human understanding and supporting decision-making across various domains. This is especially pertinent in genomics, where identifying phenotypes from clinical narratives is essential for enhancing diagnostic precision and enabling personalized medicine. While current methods perform well in recognizing explicitly stated phenotypes, they often struggle to capture implicit or nuanced representations.

In this paper, we introduce a novel workflow that integrates Large Language Models (LLMs) with Retrieval-Augmented Generation (RAG) and hierarchical reranking, utilizing hyperbolic embeddings trained on the Human Phenotype Ontology (HPO). Furthermore, we contend that conventional evaluation frameworks relying on exact string matching are insufficient for comprehensive performance assessment, as they fail to account for the hierarchical structure inherent to the target ontology. To address this, we propose new evaluation metrics that leverage the hierarchical relationships within HPO.

Our experiments on benchmark datasets, including a newly curated, challenging dataset (CHU-50), demonstrate the effectiveness of our approach, yielding substantial improvements in ranking accuracy and overall performance.

1 Introduction

The extraction of phenotypes from clinical notes is fundamental to modern diagnostic workflows. Phenotypes, as observable traits linking clinical observations to genetic conditions, play a crucial role in diagnosis, treatment planning, and biomedical research. Although recent advances in Natural Language Processing have enabled significant progress in this area, notable challenges remain, particularly in identifying implicit phenotypes that are not explicitly mentioned but can be inferred from context.

Current approaches Feng et al. (2023); Luo et al. (2021); Arbabi et al. (2019) frequently rely on flat embedding spaces, which are inadequate for modeling the hierarchical relationships intrinsic to phenotypic ontologies such as the Human Phenotype Ontology (HPO) Robinson et al. (2008). Furthermore, retrieval-based systems are often constrained by their reliance on exact matches or shallow semantic representations. We also argue that existing evaluation metrics widely used in the field (Groza et al., 2024) present further limitations: in practice, clinicians may interpret phenotype mentions differently, as no individual possesses exhaustive knowledge of HPO or uses it in a uniform manner. Consequently, a single reference can yield multiple, equally valid annotations. This underscores the importance of considering hierarchical relationships, such as treating a parent term of a target phenotype as correct, albeit less specific.

In this paper, we propose to address the following research questions:

i. To what extent can hyperbolic models capture the hierarchical structure of the HPO ontology?

ii. What is the true performance of Retrieval-Augmented Generation (RAG) Lewis et al. (2020) for phenotype candidate retrieval?

iii. What is the added value of hyperbolic embeddings in the retrieval and reranking process?

iv. How does incorporating ontology hierarchy into evaluation metrics affect performance assessment?

To tackle these challenges, we propose a novel workflow that integrates Large Language Models (LLMs) for span identification, RAG for candidate generation, and hyperbolic embeddings for hierarchical reranking. Furthermore, we introduce a hierarchy-aware evaluation metric designed to

more fairly assess ontology-based entity extraction from text. By leveraging the hierarchical structure of HPO, our approach aims to enhance the relevance of phenotype extraction from clinical narratives.

2 Background

The introduction of ontologies such as the HPO has provided a structured framework for organizing phenotypic information and has become the primary target for entity linking in this domain. Early work Aronson and Lang (2010); Jonquet et al. (2009); Deisseroth et al. (2019), utilized rule-based heuristics, while more recent studies have adopted transformer-based architectures to extract phenotype mentions directly from text Feng et al. (2023); Yang et al. (2024). Although improvements have been effective with such approaches, they remain complex and often struggle when phenotype references are implicit Baddour et al. (2024) . Emerging paradigms such as RAG Lewis et al. (2020) offer a promising avenue for addressing some of these challenges by efficiently narrowing the candidate space. However, RAG has not yet been widely adopted in phenotype extraction pipelines, and its performance in this context remains underexplored.

While ontologies facilitate annotation and retrieval, their hierarchical complexity poses significant challenges for NLP systems. Nickel and Kiela (2017) highlighted the limitations of flat embedding spaces in adequately representing such hierarchical structures. Related works (Sala et al., 2018; Sinha et al., 2024; Tifrea et al., 2018) proposed to train hyperbolic embeddings that provide a compelling alternative, as hyperbolic spaces are well-suited for modeling hierarchical relationships, allowing embeddings to more accurately reflect the subsumption structure inherent in ontologies.

The motivation behind our proposed workflow stems from recognizing significant limitations in current phenotype extraction systems. While classical RAG approaches are effective at retrieving candidates based on general semantic similarity, they fall short in capturing the hierarchical relationships and intricate dependencies inherent in ontologies such as HPO. This limitation becomes even more pronounced when dealing with implicit phenotypes not explicitly stated in clinical text, where leveraging ontological relationships can be crucial for accurate identification and resolution.

By integrating hyperbolic embeddings (which naturally encode hierarchical structures) with a reranking mechanism, our workflow bridges the gap between general semantic relevance and ontological hierarchy. This dual approach ensures not only accurate retrieval of phenotypes but also a ranking that reflects their hierarchical significance, providing a comprehensive solution to the limitations of current methods.

3 Proposed Workflow

The preliminary processing phase involves manual annotation of clinical reports (using existing annotations for the ID-68 dataset and newly created annotations for the CHU-50 dataset) as well as fine-tuning the hyperbolic model on the HPO ontology (the training procedure is detailed in Section II.B). The high-level architecture of HyperRAG is illustrated in Figure 1. Given clinical reports and their annotations, the process consists of four main steps: span detection using an LLM, candidate retrieval with RAG, reranking of candidates, and evaluation with both standard and ontology-aware metrics.

3.1 Span Identification

We begin by leveraging a pretrained Large Language Model (LLM) to identify phenotype spans within clinical text. This unsupervised approach is particularly effective for capturing implicit mentions that may be overlooked by traditional methods. Notably, Baddour et al. (2024) demonstrated that employing an LLM as a span detector outperforms the biomedical Stanza pipeline (Zhang et al., 2021). For consistency and comprehensive coverage, we utilized the same ChatGPT-3.5 (OpenAI, 2023) model employed in their work for this step.

3.2 Retrieval-Augmented Generation (RAG)

A classical embeddings model (*all-MiniLM-L12-v2*, Wang, 2020) is used to compute dense embeddings for the identified spans. Alternatively, a fine-tuned hyperbolic model *HiT-MiniLM-L12-HPO* (fine-tuned on HPO from *all-MiniLM-L12-v2*) is used. *Top-k* phenotype candidates are retrieved from the HPO ontology based on *cosine similarity* (euclidean model) or *hyperbolic distance* (hyperbolic model). We set *k=30* to substantially reduce the candidate space while still allowing for meaningful reranking improvements.

For candidate retrieval, FAISS (Douze et al., 2024) is employed as the vector store and Top-*k* retriever for the Euclidean model. In contrast,

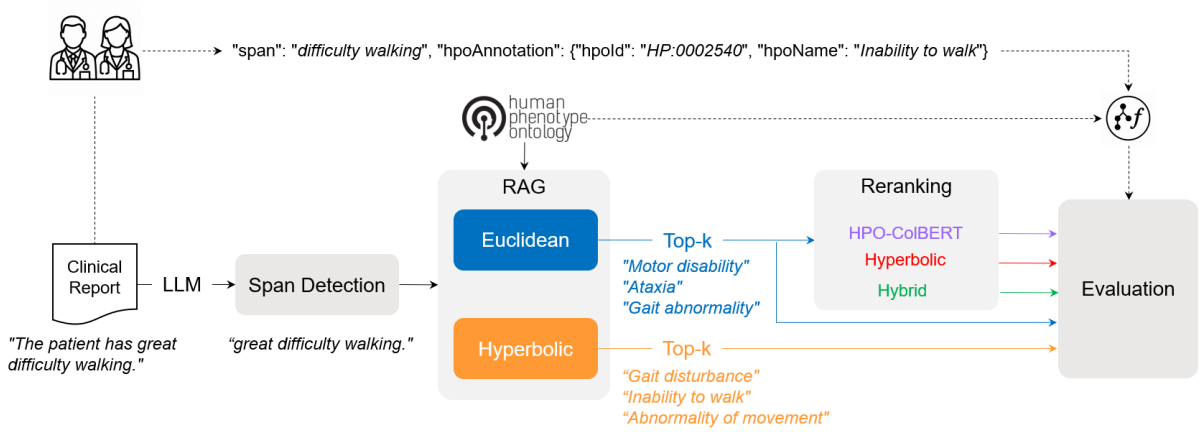

Figure 1: General Workflow

the hyperbolic model utilizes a dedicated vector index and retrieval mechanism implemented in Torch (Paszke et al., 2019). To ensure consistency in distance measurements across experiments, we normalize the hyperbolic distances in the Poincaré ball using a global normalization strategy (1):

$$\hat{d}_{\mathbb{H}}(u,v) = \frac{d_{\mathbb{H}}(u,v)}{\max_{p,q \in \text{HPO}} d_{\mathbb{H}}(p,q)} \quad (1)$$

Where:

$d_{\mathbb{H}}(u,v)$ is the hyperbolic distance between terms u and v in the hyperbolic space $\mathbb{H}$

$\hat{d}_{\mathbb{H}}$ is the normalized hyperbolic distance

$\max_{p,q \in \text{HPO}} d_{\mathbb{H}}(p,q)$ is the maximum hyperbolic distance between any two terms in the HPO ontology

3.3 Reranking

For each span, the Top-30 candidates retrieved by Euclidean-RAG are reranked using two families of methods: a classical state-of-the-art baseline and hyperbolic-based.

Late-interaction reranking

To provide a strong classical baseline, we fine-tuned a late-interaction model ColBERTv2 (Santhanam et al., 2021) for reranking. While cross-encoder models are highly effective for reranking tasks, they are computationally intensive and may be less suitable for incorporating soft signals such as distance-based scores. Late-interaction models, such as ColBERTv2, offer a compelling compromise between cross-encoders and bi-encoders by retaining token-level embeddings and applying a late matching function. This approach preserves fine-grained information that might otherwise be lost during token pooling, as in bi-encoder models. Given the short spans and specific target labels in our setting, late-interaction models are particularly well-suited for reranking.

Hyperbolic-based reranking

- **Full hyperbolic reranking:** Both the input span and the Top-k candidates from the Euclidean RAG are embedded in hyperbolic space. Candidates are then reordered based on their normalized hyperbolic distances to the input span.
- **Hybrid reranking:** This approach combines the cosine similarity between Euclidean embeddings and the hyperbolic distance between hyperbolic embeddings using a weighted sum. Cosine similarity emphasizes semantic closeness, while hyperbolic distance prioritizes candidates with closer hierarchical relationships to the input span.

$$S_{\text{hybrid}} = \gamma \cdot S_{\cos} - (1-\gamma) \cdot \hat{d}_{\mathbb{H}} \quad (2)$$

where:

S_{hybrid} represents the hybrid scores

$S_{\cos}$ represents the cosine similarities

$\hat{d}_{\mathbb{H}}$ represents the normalized hyperbolic distances

γ is the weighting parameter between the two metrics. We set $\gamma = 0.5$ to have a balance influence of both models.

Synonyms in the RAG output are mapped to their original HPO terms using a precomputed synonym-to-ID mapping. This ensures consistency in distance calculations throughout the workflow.

4 Dataset

4.1 Ontologies

The Human Phenotype Ontology (HPO) serves as the foundation for our hierarchical embeddings. HPO is a comprehensive ontology encompassing over 19,000 phenotypic terms, each organized within a rich hierarchical structure. The ontology also incorporates synonyms to account for alternative term representations, enhancing its coverage and utility.

Additionally, we leverage the SNOMED (El-Sappagh et al., 2018) ontology indirectly through a pretrained hyperbolic model. This allows us to assess the relative benefits of utilizing a broad, general-purpose medical ontology (SNOMED) in comparison to a highly specialized ontology (HPO) for phenotype extraction tasks.

4.2 Hyperbolic Training Data

Hyperbolic embeddings were trained on the HPO ontology using Hierarchy Transformers (He et al., 2024) to effectively capture hierarchical relationships. These embeddings encode both parent-child and sibling relationships, enabling more nuanced phenotypes reranking. To construct the training data for HPO hyperbolic embeddings, we first extracted hierarchical relationships from the HPO OWL file using DeepOnto and the ELK reasoner. Following the methodology of (He et al., 2024), we generated a dataset of pairs ({child, parent, label}, where the label is a binary indicator of a positive or negative example) and triplets ({child, parent, negative}, where the negative term is not a parent of the child). Random negative sampling was employed in this implementation, though hard negative sampling remains an alternative. Given that most HPO phenotypes are associated with multiple synonyms, we augmented the dataset by including all possible synonym combinations within each pair and triplet. This augmentation enhances the robustness of the resulting embeddings to varied term formulations. To prevent excessive class imbalance, we applied a filtering strategy, limiting each synonym to a maximum of five occurrences. The final pairs and triplets datasets were then used to train the hyperbolic embeddings.

Training parameters are provided in the supplementary materials.

4.3 Late-interaction Training Data

We fine-tuned the ColBERTv2 model on pairs of the form {span, HPO label, score}, where the score represents a similarity measure. To construct a comprehensive training dataset, we used ChatGPT-4o-mini to generate 10 clinical report sentences for each HPO term in the ontology. To ensure diversity and representativeness, we specified requirements for each batch of 10 sentences (e.g., at least two sentences should be implicit, up to two should include measurements, etc.). For this iteration, we excluded cases where a sentence refers to multiple phenotypes. For each generated sentence, we further prompted ChatGPT-4o-mini to extract the most precise span capturing the clinical observation of the target phenotype. This process resulted in the *HPO_HR_sentences_spans* dataset, comprising over 200,000 clinical sentences and corresponding spans, covering the entire set of HPO terms.

To improve data quality, we applied heuristic filtering to remove lower-quality spans, yielding 91,760 spans (with 2,167 unique spans filtered out). For scoring, we leveraged the trained hyperbolic model: positive (span, label) pairs from the generated dataset were assigned a score of 1, while negative pairs were created by pairing spans with other phenotypes and assigning scores based on the normalized hyperbolic distance to the target phenotype. Both hard negatives (phenotypes within the same branch, up to three hops away) and easy negatives (phenotypes outside the target branch) were included. The final training set consists of 510,371 pairs.

4.4 Evaluation Dataset

We evaluate our workflow using two datasets: **ID-68:** A widely used benchmark for phenotype extraction (Anazi et al., 2017). **CHU-50:** An internal dataset containing anonymized clinical notes from Rennes Hospital with a high proportion of implicit annotations. Results are compared to PhenoBERT, the most advanced open-source state-of-the-art solution available.

5 Evaluation

We first evaluated the hyperbolic model independently, prior to conducting the main phenotype extraction experiments.

5.1 Hyperbolic Inner Evaluation

To assess the consistency of the hyperbolic model, we compared its normalized distance metrics with those of the baseline Euclidean model. Specifically, we examined one-hop and multi-hop distances to evaluate the model's ability to capture hierarchical relationships, as well as distances between synonyms and negative pairs to determine whether semantic consistency is preserved.

Additionally, we introduce a *hierarchical representation power* plot to visualize the model's capacity to encode hierarchy while maintaining semantic coherence. This radar chart displays the average distances for one-hop, multi-hop, and synonym pairs, alongside the inverse average distance for negative pairs. This visualization enables us to assess whether the embedding space has been structured as intended.

5.2 Phenotypes Linking Evaluation

In practice, generating a comprehensive list of phenotypes for each patient is crucial for accurate diagnosis, making recall-based metrics (recall@k and miss_rate@k) the primary focus. While Top-1 precision is reported for comparison with existing methods, it can be biased by clinician habits and is less informative at higher ranks. To further assess ranking quality, we include Mean Reciprocal Rank (MRR) and Normalized Discounted Cumulative Gain (NDCG).

However, these traditional metrics are limited when based solely on exact matches, which is the prevailing evaluation paradigm in current solutions. In practice, a parent term of the target phenotype often conveys relevant information, even if it is less specific, and predictions involving descendants or related terms should not be considered entirely incorrect.

To address this limitation, we introduce a novel hierarchical evaluation framework that leverages the structure of HPO to weight candidate scores according to their proximity to the ground truth. These *relationship scores* are computed based on the specific type of relationship between the candidate $\mathcal{C}$ and the target phenotype $\mathcal{T}$.

Direct relationships

$$w_{direct}(\mathcal{C},\mathcal{T}) = \begin{cases} \frac{\alpha}{p\times(1+|d|)}, & p>0 \\ 1, & p=0 \end{cases} \quad (4)$$

where:

α is a constant factor (set to 1.6 in our experiments).

p is the number of ancestors/children between $\mathcal{C}$ and $\mathcal{T}$.

d is the distance between $\mathcal{C}$ and $\mathcal{T}$.

Indirect relationship

$$w_{indirect} = \frac{\beta}{c\times(1+d_l)} \quad (5)$$

where:

β is a constant factor. (set to 1.0 in our experiments).

c is the number of immediate children of the most specific common ancestor between $\mathcal{C}$ and $\mathcal{T}$.

d_l is the distance between $\mathcal{C}$ and the farthest HPO leaf.

By combining absolute distances with the cardinality of surrounding phenotypes, these functions effectively characterize the strength of relationships between HPO terms, balancing both proximity and semantic relevance. Throughout this paper, the term *weighted* metrics refers to evaluation metrics that incorporate these hierarchical weightings.

In addition, we introduce specific metrics to assess how well the models respect the ontology's structure: the average number of hops between each candidate and the target phenotype; the average branch coverage, defined as the proportion of candidates within the same branch as the target; and the distribution of relationship types by position, measuring the proportions of exact matches, ancestors, descendants, cousins, or candidates with no direct path to the target. We also report the proportion of close candidates, defined as those with a *relationship score* above a specified threshold.

6 Results

6.1 Hyperbolic consistency

Figure 2. presents the distributions of one-hop and multi-hop distances for both the Euclidean model and the fine-tuned hyperbolic model. The distributions for the hyperbolic model are notably narrower and exhibit lower means, particularly for multi-hop distances, indicating a more faithful representation of the ontology's hierarchical structure.

Furthermore, the resulting hyperbolic model preserves the semantic structure of the base model,

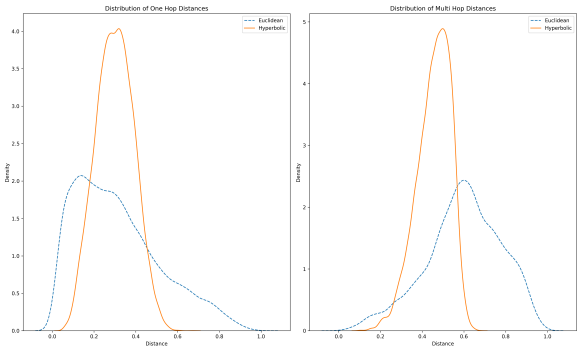

Figure 2: Euclidean vs Hyperbolic Distances Distribution

as illustrated in Figure 3. Although the average distance between negative pairs is slightly reduced, these pairs remain well separated from positive examples. Notably, synonyms within the HPO are now positioned closer together, and multi-hop phenotypes are significantly closer than in the Euclidean embedding space, reflecting improved hierarchical modeling. In contrast, one-hop phenotypes are only marginally closer, which is expected given the typically strong semantic similarity between such terms (e.g.: *Iris coloboma* is semantically closer to its one-hop parent *Coloboma* than the 2-hops *Abnormal eye morphology*).

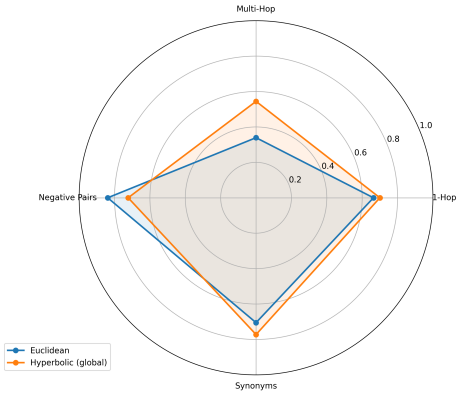

Figure 3: Semantic and Hierarchical Representation Power

6.2 Phenotypes Linking

Retrieval

As shown in Figure 4, the hyperbolic RAG model underperforms compared to other approaches, with recall decreasing when hyperbolic reranking is applied to Euclidean RAG candidates. Hybrid reranking, however, improves recall from k=5 onwards, and late-interaction reranking becomes effective from k=15, though it does not

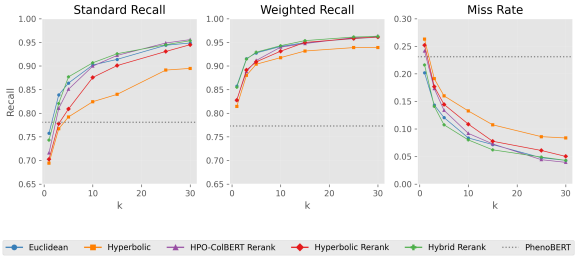

Figure 4: Recall and Miss Rate (ID-68)

surpass the hybrid method. The Euclidean model performs strongly, confirming RAG retrieval as a robust baseline. The logarithmic shape of the recall and miss rate curves indicates that all models rank correct candidates highly.

Weighted recall shows a similar pattern, with ontology-aware metrics especially benefiting hyperbolic approaches and narrowing the performance gap. Both Euclidean and hybrid reranking outperform previous SOTA recall from k=3 onwards, and set new SOTA at k=1 in the weighted setting (+9), with gains up to +18 at k=15. Hybrid reranking also achieves the lowest miss rate, reducing misses by 17 at k=15. The precision at k=1 reaches 0.857 for the Euclidean model.

Evaluation on the CHU-50 dataset (Figure 5) yields similar trends, with all models demonstrating a marked improvement over the state-of-the-art. Specifically, all approaches outperform PhenoBERT across all three metrics starting from (k=1), achieving a +23 increase in recall and an 18-point reduction in miss rate. As previously noted by Baddour et al. (2024), this is expected, as PhenoBERT exhibits limitations in handling more implicit phenotype references.

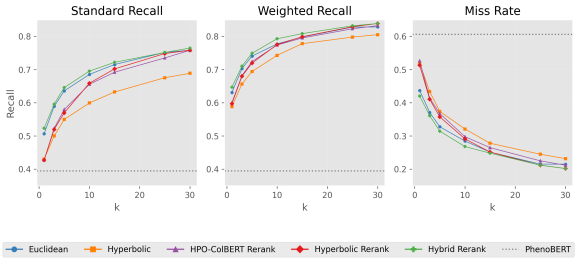

Figure 5: Recall and Miss Rate (CHU-50)

Finally, we evaluated the same metrics using a hyperbolic model fine-tuned on the SNOMED ontology instead of HPO. As shown in Figure 6, the SNOMED-based hyperbolic model underperforms compared to the HPO-based hyperbolic model, which is expected given that the target phenotypes

are defined within the HPO ontology. Interestingly, the hybrid approach exhibits a slight improvement on the ID-68 dataset. This counterintuitive result may be attributed to the fact that, for the most challenging text spans, classical embeddings outperform hyperbolic ones, and the cosine similarity component can dominate the hybrid scoring when the hyperbolic model is less effective.

Conversely, the hybrid model demonstrates reduced accuracy with SNOMED on the CHU-50 dataset.

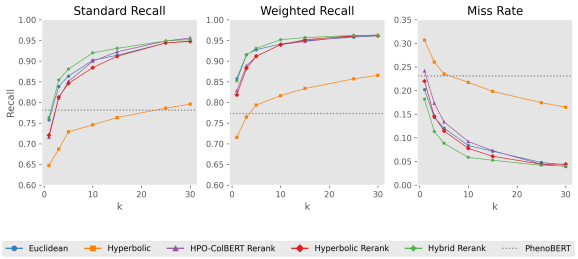

Figure 6: Recall and Miss Rate with Snomed (ID-68)

Ranking

Tables 1 and 2 present weighted MRR and NDCG results for ID-68 and CHU-50. On ID-68, the Euclidean and Hybrid Rerank models achieve the highest MRR (e.g., 0.857 at k=1), indicating top-ranked correct phenotypes. The Hyperbolic model performs slightly lower, but the gap narrows with hierarchy-aware metrics, highlighting its strength in capturing ontological relationships. NDCG scores are also high across all models, with Euclidean and Hybrid Rerank exceeding 0.94 at k=1. As k increases, both metrics decrease slightly, but Hybrid Rerank consistently maintains strong performance.

Table 1: Weighted Metrics by Model and k (ID-68)

Weighted MRR					
Model	$k=1$	$k=3$	$k=5$	$k=10$	$k=15$
Euclidean	**0.857**	**0.882**	**0.884**	**0.885**	**0.885**
Hyperbolic	0.814	0.841	0.845	0.847	0.848
HPO-ColBERT Rerank	0.828	0.851	0.855	0.858	0.858
Hyperbolic Rerank	0.828	0.853	0.857	0.859	0.860
Hybrid Rerank	**0.855**	**0.881**	**0.883**	**0.885**	**0.885**
Weighted NDCG					
Model	$k=1$	$k=3$	$k=5$	$k=10$	$k=15$
Euclidean	**0.949**	**0.936**	**0.923**	**0.901**	0.883
Hyperbolic	0.932	0.922	0.902	0.877	0.861
HPO-ColBERT Rerank	0.931	0.928	0.912	0.886	0.870
Hyperbolic Rerank	0.944	0.927	0.912	0.884	0.870
Hybrid Rerank	0.943	**0.936**	**0.924**	**0.904**	**0.891**

On the more challenging CHU-50 dataset, which contains a higher proportion of implicit phenotype mentions, all models exhibited lower MRR and NDCG scores compared to ID-68. However,

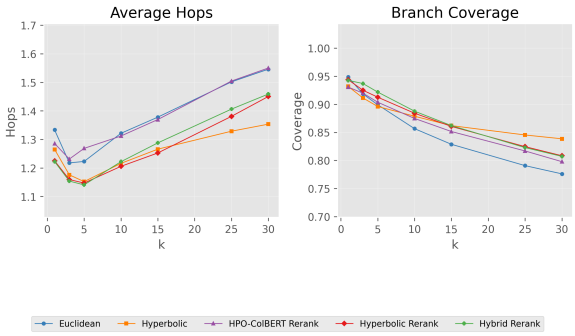

Figure 7: Ontological Structure Metrics (ID-68)

the Hybrid Rerank model still outperforms others, with an MRR of 0.647 and NDCG of 0.868 at k=1. These results indicate that the hybrid approach, which combines semantic similarity from Euclidean embeddings with hierarchical proximity from hyperbolic embeddings, is particularly effective in ranking the most relevant phenotypes at the top, even in complex, real-world clinical text.

Table 2: Weighted Metrics by Model and k for CHU-50

Weighted MRR					
Model	$k=1$	$k=3$	$k=5$	$k=10$	$k=15$
Euclidean	0.631	0.659	0.667	0.670	0.671
Hyperbolic	0.588	0.613	0.619	0.625	0.627
Hyperbolic Rerank	0.597	0.629	0.637	0.643	0.644
Hybrid Rerank	**0.647**	**0.669**	**0.676**	**0.680**	**0.681**
Weighted NDCG					
Model	$k=1$	$k=3$	$k=5$	$k=10$	$k=15$
Euclidean	0.841	**0.853**	**0.832**	**0.807**	0.792
Hyperbolic	0.834	0.818	0.796	0.772	0.751
Hyperbolic Rerank	0.851	0.841	0.815	0.787	0.779
Hybrid Rerank	**0.868**	0.848	**0.835**	**0.806**	**0.800**

Overall, the consistently strong MRR and NDCG scores for the Hybrid Rerank model confirm that combining semantic and hierarchical signals yields superior candidate ranking. Hierarchy-aware weighted metrics further demonstrate the value of hyperbolic embeddings in capturing nuanced ontological relationships, especially when exact matches are unavailable but related terms remain clinically relevant.

Ontology-based Metrics

Analyzing the number of hops between candidates and the target phenotype, as well as branch coverage (Figure 7), offers further insight into model performance. While hyperbolic-based models may not always outperform Euclidean models at top ranks, they show greater robustness as the candidate list grows, maintaining lower average hops and higher branch coverage.

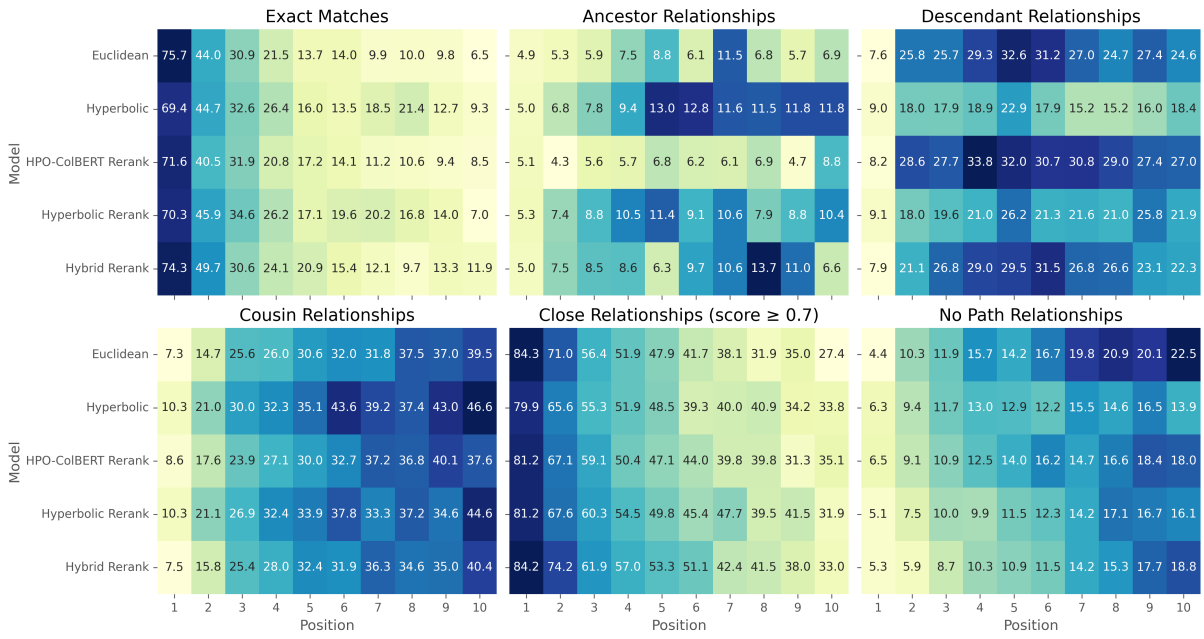

Figure 8: % of Relationship Types by Position (ID-68)

This observation is further supported by a detailed analysis of ontological relationships. Figure 8 shows a high percentage of exact matches at Top-1, confirming the RAG pipeline's effectiveness. However, deeper analysis reveals important distinctions between modeling approaches. Hyperbolic models (both raw output and reranking) exhibit significantly higher proportions of ancestor and cousin relationships, while showing fewer descendant relationships compared to Euclidean or HPO-ColBERT models. This pattern strongly suggests that hyperbolic approaches better capture the hierarchical structure of the HPO ontology in both vertical and horizontal dimensions. The tendency to "move upward" in the hierarchy toward more general terms rather than "downward" toward more specific ones aligns with theoretical expectations of hyperbolic geometry, where distances increase exponentially with depth in the hierarchy.

Notably, hyperbolic models maintain semantic relevance at higher ranks, preserving close relationships and yielding fewer unrelated candidates as k increases. This semantic consistency at higher ranks has important implications for clinical applications, as it reduces the risk of missing relevant phenotypes (false negatives) when examining a broader set of candidates.

The hybrid reranking approach combines the strengths of both geometries, achieving strong exact matching at top positions and semantic coherence at higher ranks. This balanced performance confirms the value of integrating both approaches for optimal phenotype retrieval in clinical settings. Similar trends are observed on the CHU-50 dataset (appendix B).

7 Conclusion

In summary, this work introduces HyperRAG, a novel pipeline that synergistically combines LLM-based span detection, retrieval-augmented generation, and hierarchical reranking using hyperbolic embeddings for phenotype linking from clinical text. Through comprehensive experiments on both benchmark and challenging real-world datasets, the approach demonstrates substantial improvements in recall, miss rate, and ranking quality, particularly when evaluated with hierarchy-aware metrics that better reflect clinical relevance. The hybrid reranking strategy, integrating both semantic and ontological signals, consistently delivers state-of-the-art performance, especially in scenarios with implicit phenotype mentions. The proposed evaluation framework and publicly released datasets further advance the field by enabling more nuanced and clinically meaningful assessment of phenotype extraction systems. Future work should focus on enhancing the semantic modeling of implicit mentions, expanding to additional ontologies, and optimizing the computational efficiency of the pipeline for broader clinical adoption.

Limitations

While our approach demonstrates notable improvements, several limitations remain. First, although hyperbolic embeddings are effective at capturing hierarchical relationships within the ontology, they may fail to fully capture implicit semantic nuances present in clinical text. This highlights the need for future work to explore multi-task training strategies, such as fine-tuning RAG embeddings with a dual loss function that combines a hyperbolic loss for hierarchy with a sentence similarity loss to better address implicit semantics.

Another limitation concerns data sparsity: our reliance on annotated datasets may restrict the generalizability of the approach to unseen or rare phenotypes. Additionally, hyperbolic distances are not always meaningful for all term pairs, suggesting that further refinement of the hyperbolic model is warranted.

Due to computational and resource constraints, we did not perform an exhaustive grid search over the parameters for hybrid reranking or the weighting schemes used in the evaluation metrics. This may have limited the optimality of our results. Finally, our experiments relied on a single embedding model; it is possible that alternative or ensemble embedding models, particularly those fine-tuned on medical data, could yield different or improved results, potentially surpassing the performance of our current hybrid reranking approach.

References

Shams Anazi, Sateesh Maddirevula, Vincenzo Salpietro, Yasmine T Asi, Saud Alsahli, Amal Alhashem, Hanan E Shamseldin, Fatema AlZahrani, Nisha Patel, Niema Ibrahim, and 1 others. 2017. Expanding the genetic heterogeneity of intellectual disability. Human genetics, 136:1419–1429.

Aryan Arbabi, David R Adams, Sanja Fidler, Michael Brudno, and 1 others. 2019. Identifying clinical terms in medical text using ontology-guided machine learning. JMIR medical informatics, 7(2):e12596.

Alan R Aronson and François-Michel Lang. 2010. An overview of metamap: historical perspective and recent advances. Journal of the American Medical Informatics Association, 17(3):229–236.

Moussa Baddour, Stéphane Paquelet, Paul Rollier, Marie De Tayrac, Olivier Dameron, and Thomas Labbé. 2024. Phenotypes extraction from text: Analysis and perspective in the llm era. In 2024 IEEE 12th International Conference on Intelligent Systems (IS), pages 1–8. IEEE.

Cole A Deisseroth, Johannes Birgmeier, Ethan E Bodle, Jennefer N Kohler, Dena R Matalon, Yelena Nazarenko, Casie A Genetti, Catherine A Brownstein, Klaus Schmitz-Abe, Kelly Schoch, and 1 others. 2019. Clinphen extracts and prioritizes patient phenotypes directly from medical records to expedite genetic disease diagnosis. Genetics in Medicine, 21(7):1585–1593.

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

Shaker El-Sappagh, Francesco Franda, Farman Ali, and Kyung-Sup Kwak. 2018. Snomed ct standard ontology based on the ontology for general medical science. BMC medical informatics and decision making, 18:1–19.

Yuhao Feng, Lei Qi, and Weidong Tian. 2023. Phenobert: A combined deep learning method for automated recognition of human phenotype ontology. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 20(2):1269–1277.

Tudor Groza, Harry Caufield, Dylan Gration, Gareth Baynam, Melissa A Haendel, Peter N Robinson, Christopher J Mungall, and Justin T Reese. 2024. An evaluation of gpt models for phenotype concept recognition. BMC Medical Informatics and Decision Making, 24(1):30.

Yuan He, Zhangdie Yuan, Jiaoyan Chen, and Ian Horrocks. 2024. Language models as hierarchy encoders. In Advances in Neural Information Processing Systems, volume 37, pages 14690–14711. Curran Associates, Inc.

Clement Jonquet, Nigam H Shah, Cherie H Youn, Mark A Musen, Chris Callendar, and Margaret-Anne Storey. 2009. Ncbo annotator: semantic annotation of biomedical data. In ISWC 2009-8th International Semantic Web Conference, Poster and Demo Session, 171.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in neural information processing systems, 33:9459–9474.

Ling Luo, Shankai Yan, Po-Ting Lai, Daniel Veltri, Andrew Oler, Sandhya Xirasagar, Rajarshi Ghosh, Morgan Similuk, Peter N Robinson, and Zhiyong Lu. 2021. Phenotagger: a hybrid method for phenotype concept recognition using human phenotype ontology. Bioinformatics, 37(13):1884–1890.

Maximillian Nickel and Douwe Kiela. 2017. Poincaré embeddings for learning hierarchical representations. Advances in neural information processing systems, 30.

OpenAI. 2023. Chatgpt (mar 23 version). Https://chat.openai.com/chat.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, and 1 others. 2019. Pytorch: An imperative style, high-performance deep learning library. In Advances in Neural Information Processing Systems, volume 32.

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.

Frederic Sala, Chris De Sa, Albert Gu, and Christopher Ré. 2018. Representation tradeoffs for hyperbolic embeddings. In International conference on machine learning, pages 4460–4469. PMLR.

Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. 2021. Colbertv2: Effective and efficient retrieval via lightweight late interaction. arXiv preprint arXiv:2112.01488.

Aditya Sinha, Siqi Zeng, Makoto Yamada, and Han Zhao. 2024. Learning structured representations with hyperbolic embeddings. Advances in Neural Information Processing Systems, 37:91220–91259.

Alexandru Tifrea, Gary Bécigneul, and Octavian-Eugen Ganea. 2018. Poincar\'e glove: Hyperbolic word embeddings. arXiv preprint arXiv:1810.06546.

Wei F. Dong L. Bao H. Yang N. Zhou M. Wang, Y. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In NeurIPS.

Jingye Yang, Cong Liu, Wendy Deng, Da Wu, Chunhua Weng, Yunyun Zhou, and Kai Wang. 2024. Enhancing phenotype recognition in clinical notes using large language models: Phenobcbert and phenogpt. Patterns, 5(1).

Yuhao Zhang, Yuhui Zhang, Peng Qi, Christopher D Manning, and Curtis P Langlotz. 2021. Biomedical and clinical english model packages for the stanza python nlp library. Journal of the American Medical Informatics Association, 28(9):1892–1899.

A Training Settings

All model training was conducted on a single RTX A3000 GPU, both to accommodate budget constraints and to reduce energy consumption for environmental considerations.

Table 3 indicates the training settings for the hyperbolic model training.

The training settings for the ColbertV2 fine-tuning on HPO is presented in Table 4.

Parameter	Value
Number of training epochs	20
Train batch size	32
Eval batch size	64
Learning rate	1e-5
Clustering loss weight	1.0
Clustering loss margin	5.0
Centripetal loss weight	1.0
Centripetal loss margin	0.5
Gradient accumulation steps	8

Table 3: Hyperbolic Training Hyperparameters.

Parameter	Value
Train batch size	8
Learning rate	1e-5
Number of training epochs	2
Max query length	32
Max document length	128
Triplet loss margin	0.3
Gradient accumulation steps	2

Table 4: ColBERTv2 Training Hyperparameters.

B CHU-50 Ontological Structure Evaluation

The average number of hops between candidates and target phenotypes is shown in Figure 9. The rank scale is up to 50 so the robustness of the hyperbolic model for higher ranks is highlighted.

Figure 10 presents the relationship types proportion for the CHU-50 dataset.

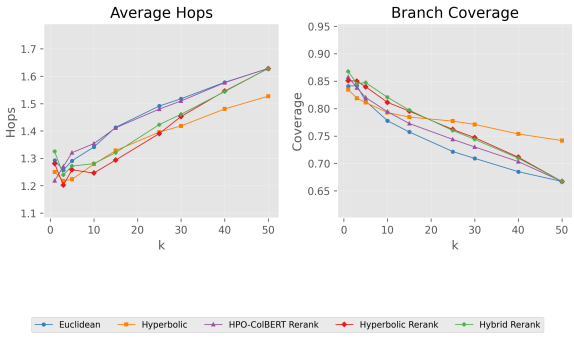

Figure 9: Average Hops and Branch Coverage (CHU-50)

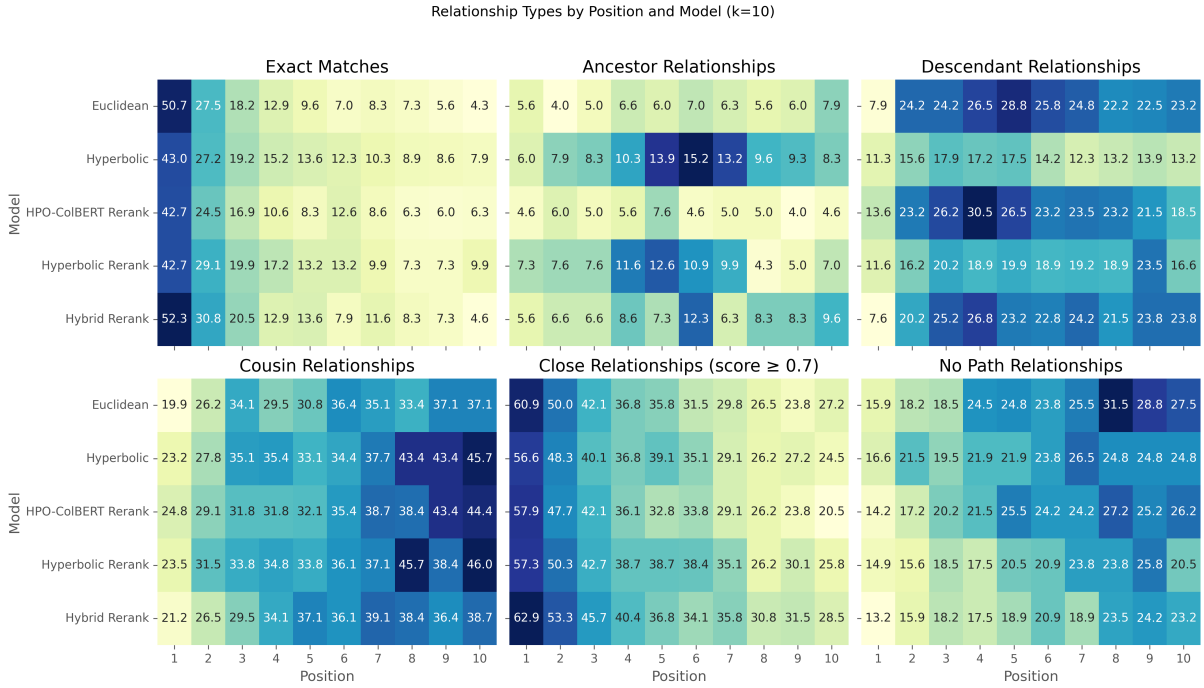

Figure 10: % of Relationship Types by Position (CHU-50)