Hierarchy-Aware Hyperbolic and Semantic Reranking for Ontology-Based Phenotype Linking

Thomas Labbé[1,2], Moussa Baddour[2], Axel Bonesteve[3,4], Paul Rollier[3], Marie de Tayrac[2,3,4], Olivier Dameron[2,5]

[1]Orange Research [2]b<>com
[3]Univ Rennes, Department of Molecular Genetics and Genomics, CHU Rennes, IGDR-UMR6290, CNRS
[4]Geeng genetic Genius [5]Univ Rennes, Inria, CNRS, IRISA - UMR 6074
thomas.labbe@gmail.com

Abstract

Extracting structured knowledge from unstructured text is a fundamental challenge in machine learning, particularly for concepts organized within complex hierarchical ontologies. In genomics, identifying phenotypes from clinical narratives is crucial for diagnostic precision, yet current methods struggle with contextual interpretation and subtle clinical descriptions. We present a hierarchy-aware workflow for ontology-based phenotype linking that combines semantic and hierarchical signals. Our approach integrates Large Language Models for span detection with retrieval and a hybrid reranking strategy using both Euclidean (semantic) and hyperbolic (hierarchical) embeddings trained on the Human Phenotype Ontology. We show that while hyperbolic embeddings alone do not outperform standard semantic retrieval, they provide complementary structural signals that improve performance over strong baselines when combined with Euclidean representations. In particular, the hybrid approach outperforms existing state-of-the-art methods and yields more hierarchically coherent predictions, especially in settings involving implicit phenotype mentions. Experiments on a public benchmark (ID-68) and a newly released clinical dataset (CHU-50), publicly released with code and data [1], highlight both performance gains and improved alignment with ontology structure. We further introduce a hierarchy-aware evaluation framework that reflects clinical relevance beyond exact-match metrics.

1 Introduction

Extracting structured knowledge from unstructured text is a central challenge in machine learning, particularly when the target concepts are organized within complex hierarchical ontologies (Figure 1). Phenotypes, as observable traits linking clinical observations to genetic conditions, play a crucial role in diagnosis, treatment planning, and biomedical research Son et al. (2018); Yuan et al. (2022); Mao et al. (2025). While recent advances in natural language processing have improved the extraction of explicitly stated phenotypes, existing systems often struggle to identify implicit or context-dependent mentions.

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). In contrast, hyperbolic geometry naturally models hierarchical (tree-like) structures Gromov (1987) and has been underexplored within this domain. 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 applies it uniformly. Consequently, a single reference can yield multiple, equally valid annotations. This variability challenges single-label prediction paradigms and motivates hierarchy-aware approaches, where related terms (e.g., ancestors of a target phenotype), are considered clinically meaningful, albeit less specific.

In this paper, our contributions are four-fold:

i. We propose a workflow for phenotype extraction from clinical text that integrates Large Language Models (LLMs) with retrieval and hierarchical reranking using hyperbolic embeddings trained on HPO.

ii. We introduce a hierarchy-aware evaluation framework that leverages the target ontology structure to provide a more nuanced and clin-

[1] https://github.com/labbeth/HyperRAG/

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 45–63

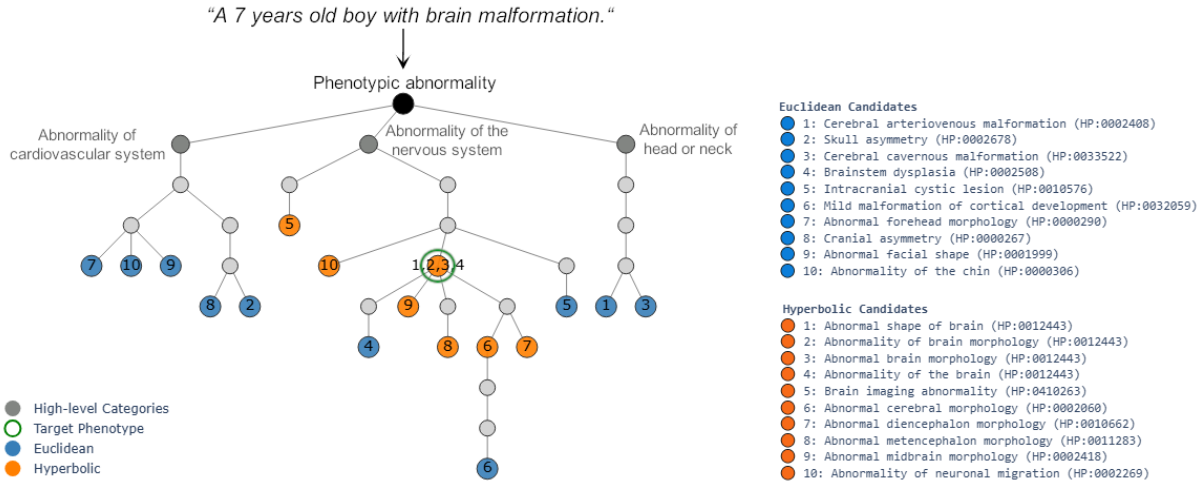

Figure 1: Ontology-based entity linking illustration. Given a clinical sentence, *Top-10* candidates from Euclidean (blue) and Hyperbolic (orange) reranking are overlaid on HPO. Hyperbolic-reranked candidates cluster within the correct branch with lower average distance to target (green), demonstrating improved hierarchical alignment.

ically relevant assessment of extraction systems.

iii. We demonstrate the effectiveness of our approach through comprehensive experiments on both benchmark and challenging real-world datasets, showing consistent improvements, particularly in scenarios with implicit entity mentions.

iv. We make all generated training and evaluation datasets publicly available through our repository, and will release trained models to support reproducibility and further research in ontology-based entity linking.

As part of our workflow validation, we also empirically confirm that hyperbolic embeddings effectively capture the hierarchical structure of HPO, supporting their use in downstream phenotype linking tasks. Overall, our ontology-aware workflow offers a robust solution for phenotype extraction from text, with the potential to enhance diagnostic accuracy and advance biomedical NLP.

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 utilized rule-based heuristics Aronson and Lang (2010); Jonquet et al. (2009); Deisseroth et al. (2019), while more recent studies have adopted transformer-based architectures to extract phenotype mentions directly from text Luo et al. (2021); 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). In particular, full LLM approaches are prone to hallucination issues in mapping phenotype labels to HPO ids Labbé et al. (2023). Emerging retrieval-based paradigms Lewis et al. (2020) offer a promising avenue for addressing some of these challenges by efficiently narrowing the candidate space and mitigating hallucinations. However, this approach 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 dense retrieval approaches Guo et al. (2024) are effective at retrieving candidates based on general semantic similarity, they fall short in capturing the hierarchical relationships and intricate depen-

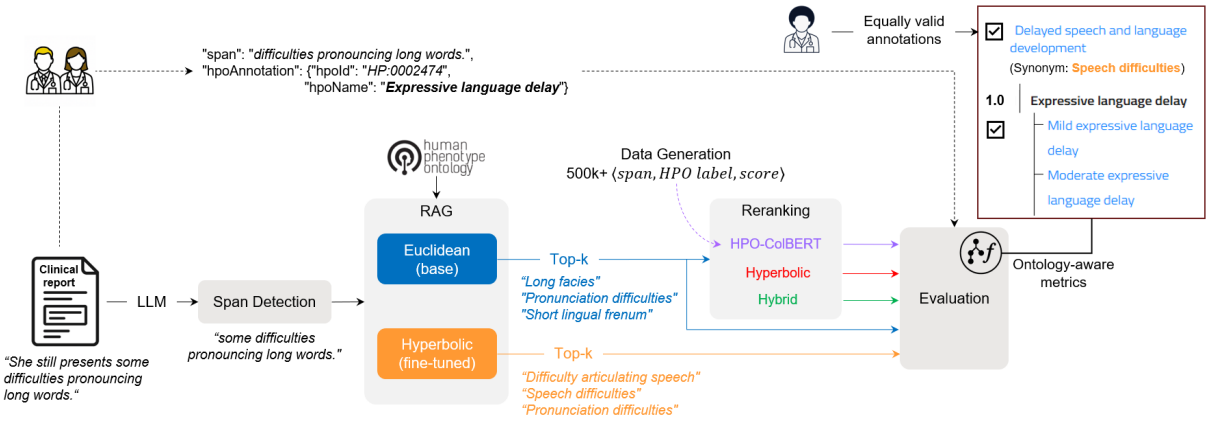

Figure 2: General Workflow. From a clinical report, a LLM first detects spans related to phenotypes. Dense retrieval with Euclidean and Hyperbolic embeddings outputs *Top-k* candidate terms, followed by several reranking strategies. Evaluation uses both clinician-provided ground truth annotations and the HPO ontology structure to account for hierarchical relationships and semantic similarity, recognizing that multiple valid references may exist.

dencies inherent in ontologies such as HPO Huang et al. (2025); Gao et al. (2023). 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 Peng et al. (2024).

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 Methodology

The proposed workflow is illustrated in Figure 2: Given clinical reports, the process consists of four main steps: span detection using an LLM, dense candidate retrieval, reranking of candidates, and evaluation with both standard and ontology-aware metrics.

3.1 Span Identification

We use a pretrained LLM to identify phenotype spans, effectively 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).

3.2 Retrieval

We compute dense span embeddings using either a base or HPO-fine-tuned hyperbolic model. *Top-k* phenotype candidates are retrieved from the HPO ontology based on *cosine similarity* (euclidean model) or normalized *hyperbolic distance* (hyperbolic model).

3.3 Reranking

For each span, the *Top-k* HPO candidates retrieved through dense Euclidean embeddings similarity are reranked using two families of methods: late-interaction and hyperbolic.

Late-interaction reranking

As a strong classical baseline, we fine-tuned a late-interaction model Santhanam et al. (2021) for reranking. While cross-encoders are effective, they are computationally demanding and less adaptable to incorporating soft signals like distance-based scores Jha et al. (2024). Late-interaction models strike a balance by retaining token-level embeddings and applying a late matching function, thus preserving fine-grained information often lost in bi-encoders. This makes them particularly suitable for our reranking task, which involves short text spans and specific target labels.

Hyperbolic reranking

- **Full hyperbolic reranking:** Both the input span and the *Top-k* candidates from the Euclidean retriever are embedded in hyperbolic space. Candidates are then reordered based on their normalized hyperbolic distances to the

input span. We include this setting primarily as an ablation to isolate the effect of purely hierarchy-driven scoring.

- **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. given an input $span$ and a set of candidates $\{C_i\}_{i=1}^{k}$, the hybrid score of a candidate C_i is computed as follows.

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

where $S_{\cos}$ is the cosine similarity between C_i and $span$, $\hat{d}_{\mathbb{H}}$ is their normalized hyperbolic distance, and γ is the weighting parameter between the two metrics (see Appendix E for ablation).

Full implementation details (including model architectures, training hyperparameters, normalization strategies, synonym mapping, and retrieval settings) are provided in Appendix A, B and D to facilitate reproducibility. Code and data are publicly available through our repository.

4 Dataset

4.1 Ontologies

The Human Phenotype Ontology (HPO) Robinson et al. (2008) serves as the foundation for our hierarchical embeddings, comprising over 19,000 hierarchically organized phenotypic terms and an extensive set of synonyms.

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

4.2 Training Data

Hyperbolic Training: Hyperbolic embeddings were trained on HPO triplets extracted from the ontology, with synonym augmentation and filtering to ensure balanced training. We used Hierarchy Transformers He et al. (2024) to effectively capture hierarchical relationships.

Late-interaction Training: We fine-tuned ColBERTv2 Santhanam et al. (2021) on triplets $\langle$*span, HPO label, score*$\rangle$, using a dataset generated by prompting ChatGPT-4o-mini to create diverse clinical sentences for each HPO term. Sentence quality was assessed by clinicians, and further validated using an LLM-as-a-judge approach. After heuristic filtering, we obtained 91,760 high-quality spans, which were used to construct positive and negative pairs, resulting in 510,371 training triplets.

Full data construction details and training parameters for both hyperbolic and late-interaction models are provided in Appendix B and D. Prompts for LLM-as-a-judge are provided in Appendix I.

4.3 Evaluation Dataset

We evaluate our workflow using two datasets. **ID-68** Anazi et al. (2017), a widely used benchmark for phenotype extraction. **CHU-50**, a dataset of 50 fully anonymized clinical notes from a partner hospital, which we publicly release trough our repository. It contains 971 phenotype annotations, approximately 30% of which are implicit. Results are compared to PhenoBERT Feng et al. (2023), 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 single-output methods (e.g. PhenoBERT), it fails to reflect the clinical relevance of alternative valid annotations. 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, p is the number of ancestors/descendants between $\mathcal{C}$ and $\mathcal{T}$, and 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, c is the number of immediate descendants of the most specific common ancestor between $\mathcal{C}$ and $\mathcal{T}$, and d_l is the distance between $\mathcal{C}$ and the farthest HPO leaf.

See Appendix E for α and β settings.

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* 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 3. 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.

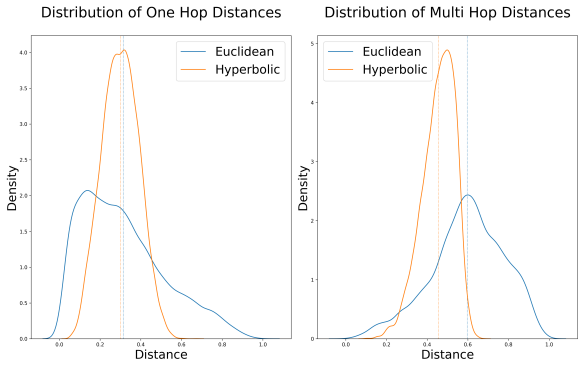

Figure 3: Euclidean (blue) vs Hyperbolic (orange) distance distributions between one-hop (left) and multi-hop (right) phenotypes in the Human Phenotypes Ontology (HPO), computed on the test set (10%). Vertical lines represent respective means.

Furthermore, the resulting hyperbolic model preserves the semantic structure of the base model, as illustrated in Figure 4. 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*).

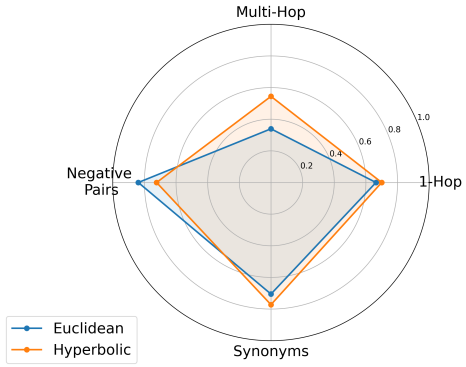

Figure 4: Semantic and Hierarchical Representation Power. Values represent normalized similarity (or inverse for negative pairs) between one-hop, multi-hop, and synonyms in Euclidean and Hyperbolic spaces.

6.2 Phenotypes Linking

For the hybrid approach, we conducted evaluations with several values of γ (reported in Appendix E). Since the performance differences were marginal, we selected $\gamma = 0.5$ for its simplicity and interpretability, recognizing that this value may not be optimal but provides a balanced influence of both models. For results reported in this section, statistical significance was assessed using paired bootstrap resampling at the report level (Appendix C).

Retrieval

As shown in Figure 5, the hyperbolic retriever model underperforms compared to other methods on both datasets, with recall further decreasing when hyperbolic reranking is applied to Euclidean retriever candidates.

This behavior reflects a trade-off between hierarchical organization and local semantic similarity: the Poincaré geometry emphasizes vertical ontological relations, while training primarily on *is-a* links and the mismatch between clinical spans and ontology-level supervision may further limit fine-grained semantic alignment, motivating the use of hyperbolic representations as a complementary rather than standalone signal.

For ID-68 (top plots), hybrid reranking improves recall from $k = 5$ onwards, and late-interaction reranking becomes effective from $k = 15$, though it does not surpass the hybrid approach. Both Euclidean and hybrid reranking outperform prior systems in recall from $k = 3$ onwards and set new state-of-the-art 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 $k = 15$.

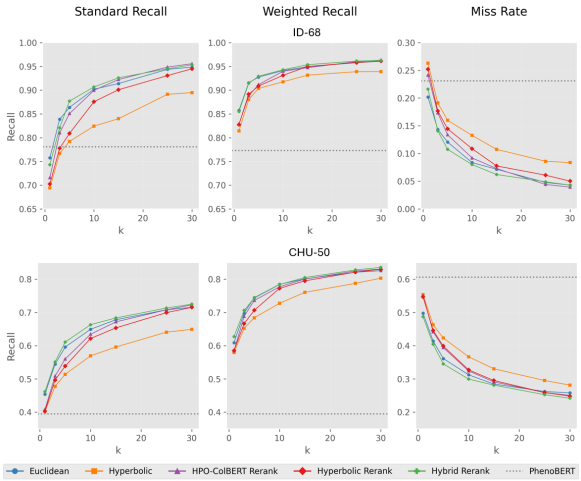

Figure 5: Recall and miss rate on the ID-68 (top) and CHU-50 (bottom) datasets. Standard (exact match) and weighted recall (considering HPO hierarchy) are shown for raw dense retrieval (Euclidean, Hyperbolic) and reranked results (HPO-ColBERT, Hyperbolic, Hybrid) across *Top-k* candidates. Miss rate indicates the proportion of ground truth phenotypes not retrieved.

On CHU-50 (bottom plots), similar trends are observed, with all models outperforming state-of-the-art PhenoBERT across all three metrics from $k = 1$, achieving a $+23$ increase in recall and an 18-point reduction in miss rate. This is consistent with previous findings Baddour et al. (2024), as PhenoBERT is less effective at capturing implicit phenotype references. Overall, the logarithmic shape of the recall and miss rate curves across both datasets indicates that all models rank correct candidates highly.

Complementary cross-ontology analyses (Appendix H) using a SNOMED-trained hyperbolic model show that, although it underperforms an HPO-trained hyperbolic model on HPO-defined targets (as expected), the SNOMED component can inject broader hierarchical cues that yield

a slight recall gain on ID-68 and a lower miss rate on CHU-50. These effects suggest that non–domain-specialized embeddings can add complementary structure in difficult cases, but the gains are modest and do not change the observation that HPO-centric retrieval and hybrid reranking remain the most effective overall, with the Euclidean baseline strong and hierarchy-aware metrics further favoring hyperbolic components.

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, 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 for ID-68

Model	$k=1$	$k=3$	$k=5$	$k=10$	$k=15$
Weighted MRR					
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					
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, the Hybrid Rerank model outperforms others for all k values.

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. 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

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

Model	$k=1$	$k=3$	$k=5$	$k=10$	$k=15$
Weighted MRR					
Euclidean	0.631	0.659	0.667	0.670	0.671
Hyperbolic	0.588	0.613	0.619	0.625	0.627
HPO-ColBERT Rerank	0.585	0.624	0.633	0.636	0.638
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					
Euclidean	0.841	**0.853**	**0.832**	**0.807**	0.792
Hyperbolic	0.834	0.818	0.796	0.772	0.751
HPO-ColBERT Rerank	0.844	0.821	0.805	0.783	0.768
Hyperbolic Rerank	0.851	0.841	0.815	0.787	0.779
Hybrid Rerank	**0.868**	0.848	**0.835**	**0.806**	**0.800**

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.

Overall, hybrid reranking yields statistically significant improvements in weighted recall@1 and weighted MRR on the CHU-50 dataset, while maintaining performance comparable to Euclidean retrieval on ID-68 (detailed statistical significance results are provided in Appendix C).

6.3 Ontology-based Metrics

While hyperbolic 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 6).

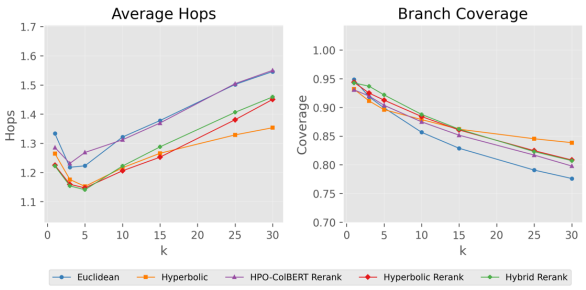

Figure 6: Ontology-based metrics on the ID-68 dataset across *Top-k* candidates. Average number of hops between the candidates and the target phenotypes (left), and proportion of candidates within the ontology branch of the target phenotypes (right).

This observation is further supported by a detailed analysis of ontological relationships (Appendix F). Figure 7 shows a high percentage of exact matches at *Top-1*, confirming the pipeline's effectiveness. However, deeper analysis reveals important distinctions between modeling approaches.

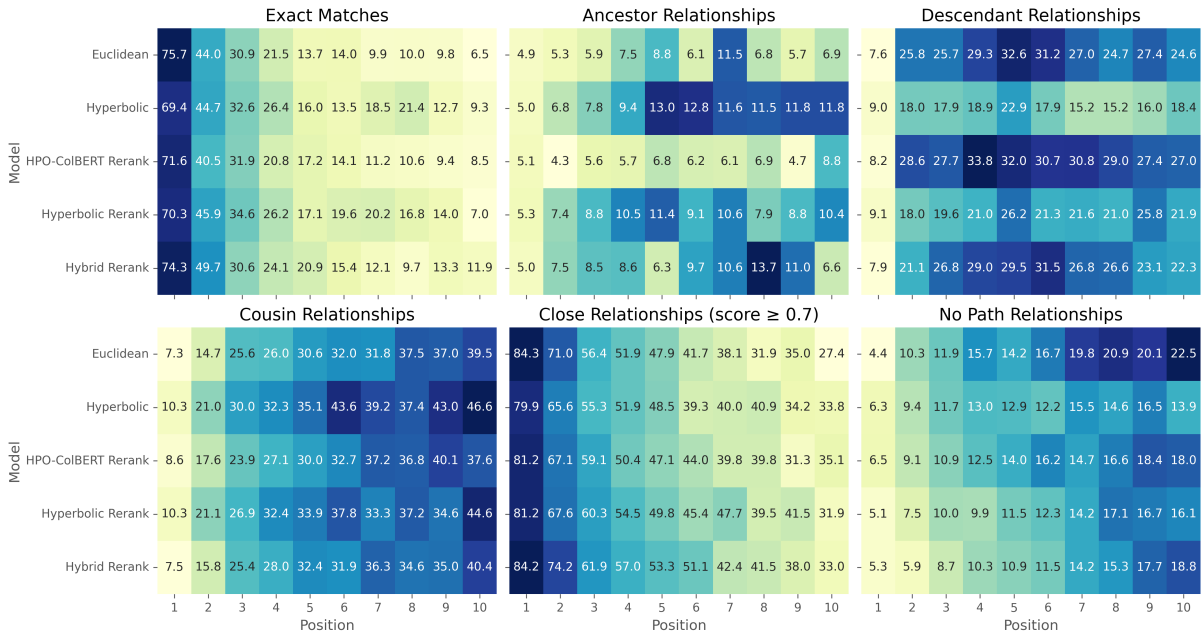

Figure 7: Distribution of relationship types by position and model for *Top-10* candidates ($k = 10$) on ID-68 dataset. The heatmaps show percentages for six relationship categories (Appendix F) between predicted and ground truth phenotypes. For closeness relationships, only those with scores above 0.7 (Equations (4) and (5)) are considered. Rows correspond to different models, and columns represent candidate positions ranked from 1 to 10. Higher values indicate stronger presence of the relationship type at the given rank and model.

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 (Figure 16 in Appendix H).

7 Conclusion

In summary, we introduce a hierarchy-aware workflow combining LLM-based span detection, dense retrieval, and hybrid reranking using hyperbolic embeddings for phenotype linking. Experiments on benchmark and real-world datasets show improvements in recall, miss rate, and ranking quality, alongside better hierarchical consistency, particularly under hierarchy-aware metrics. The hybrid reranking strategy, integrating semantic and ontological signals, consistently outperforms prior systems, especially for implicit phenotype mentions. These results highlight the complementary roles of Euclidean and hyperbolic embeddings, capturing local semantic similarity and global ontology structure, and yielding more clinically meaningful and interpretable predictions. Finally, the proposed evaluation framework and released datasets support more nuanced and clinically relevant assessment of phenotype extraction systems. This approach can be extended to other domains involving hierarchical ontologies.

Limitations

While our approach demonstrates notable improvements, several limitations remain. Although hyperbolic embeddings effectively capture hierarchical relationships within the ontology, they may fail to fully represent implicit semantic nuances in clinical text. Future work could explore multi-task training strategies jointly modeling hierarchy and semantic similarity.

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

Although CHU-50 contains a substantial proportion of implicit phenotype mentions, the current version of the dataset does not explicitly distinguish implicit and explicit annotations. Consequently, evaluations are aggregated across all mention types. Future versions of the dataset will include dedicated annotation flags to enable finer-grained analyses of implicit phenotype normalization performance.

Regarding recall performance, our analysis did not reveal clear patterns explaining why SNOMED embeddings occasionally outperform HPO embeddings. While broader-domain supervision may provide complementary signals in some cases, SNOMED embeddings underperform HPO-based models on the CHU-50 dataset, indicating that domain-specific hierarchical information remains crucial for complex clinical text.

More generally, the effectiveness of hyperbolic embeddings depends on the underlying ontology structure. Ontologies with complex or transversal relationships may limit the benefits of hyperbolic geometry, potentially requiring alternative training strategies or geometric formulations. In this respect, extending our workflow to other hierarchical coding systems such as the International Classification of Diseases (ICD) Blanco et al. (2020); Zhou et al. (2021) would require additional investigation, as its granularity, structural conventions, and limited mappings to HPO Tan et al. (2024) may constrain direct transfer.

Our approach also introduces additional system components compared to simpler pipelines such as standalone LLM inference or NER+EL models. However, most of the added cost is incurred offline, as hyperbolic embeddings are trained once and reranking relies on precomputed representations without additional LLM calls at inference time.

Finally, due to computational constraints, we did not perform an exhaustive search over hybrid reranking parameters (although we explored different values of γ in Appendix E) or evaluation weighting schemes, nor did we investigate alternative embedding backbones, particularly those fine-tuned on medical or clinical data, which could further improve performance.

Ethical Considerations

The CHU-50 dataset introduced in this work consists of clinical notes with phenotype annotations and is used exclusively for evaluation purposes. No real clinical data are used for model training. All notes were manually generated by clinicians based on real clinical reports, but the content is fully original and anonymized. No personal data are present in the dataset, and care was taken to ensure that no combination of symptoms could indirectly identify any individual. This design ensures compliance with ethical standards and protects patient privacy.

Additionally, the proposed workflow relies in part on probabilistic components, including large language models used for span detection. These outputs should be interpreted with caution. If used in a medical context, all results must be reviewed and validated by qualified healthcare professionals, and the system should not be used for autonomous clinical decision-making.

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.

Alberto Blanco, Alicia Pérez, and Arantza Casillas. 2020. Extreme multi-label icd classification: Sensitivity to hospital service and time. IEEE Access, 8:183534–183545.

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.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yixin Dai, Jiawei Sun, Haofen Wang, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. arXiv preprint arXiv:2312.10997, 2(1).

Mikhael Gromov. 1987. Hyperbolic groups. In Essays in group theory, pages 75–263. Springer.

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.

Zirui Guo, Lianghao Xia, Yanhua Yu, Tu Ao, and Chao Huang. 2024. Lightrag: Simple and fast retrieval-augmented generation.

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.

Haoyu Huang, Yongfeng Huang, Junjie Yang, Zhenyu Pan, Yongqiang Chen, Kaili Ma, Hongzhi Chen, and James Cheng. 2025. Retrieval-augmented generation with hierarchical knowledge. arXiv preprint arXiv:2503.10150.

Rohan Jha, Bo Wang, Michael Günther, Georgios Mastrapas, Saba Sturua, Isabelle Mohr, Andreas Koukounas, Mohammad Kalim Akram, Nan Wang, and Han Xiao. 2024. Jina-colbert-v2: A general-purpose multilingual late interaction retriever. arXiv preprint arXiv:2408.16672.

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.

Thomas Labbé, Pierre Castel, Jean-Michel Sanner, and Majd Saleh. 2023. Chatgpt for phenotypes extraction: one model to rule them all? In 2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), pages 1–4. IEEE.

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.

Xiaohao Mao, Yu Huang, Ye Jin, Lun Wang, Xuanzhong Chen, Honghong Liu, Xinglin Yang, Haopeng Xu, Xiaodong Luan, Ying Xiao, and 1 others. 2025. A phenotype-based ai pipeline outperforms human experts in differentially diagnosing rare diseases using ehrs. npj Digital Medicine, 8(1):68.

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.

Boci Peng, Yun Zhu, Yongchao Liu, Xiaohe Bo, Haizhou Shi, Chuntao Hong, Yan Zhang, and Siliang Tang. 2024. Graph retrieval-augmented generation: A survey. arXiv preprint arXiv:2408.08921.

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.

Jung Hoon Son, Gangcai Xie, Chi Yuan, Lyudmila Ena, Ziran Li, Andrew Goldstein, Lulin Huang, Liwei Wang, Feichen Shen, Hongfang Liu, and 1 others. 2018. Deep phenotyping on electronic health records facilitates genetic diagnosis by clinical exomes. The American Journal of Human Genetics, 103(1):58–73.

Amelia LM Tan, Rafael S Gonçalves, William Yuan, Gabriel A Brat, Robert Gentleman, Isaac S Kohane, Aaron J Masino, Adeline Makoudjou, Adem Albayrak, Alba Gutiérrez-Sacristán, and 1 others. 2024. Implications of mappings between international classification of diseases clinical diagnosis codes and human phenotype ontology terms. JAMIA open, 7(4).

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).

Xiao Yuan, Jing Wang, Bing Dai, Yanfang Sun, Keke Zhang, Fangfang Chen, Qian Peng, Yixuan Huang, Xinlei Zhang, Junru Chen, and 1 others. 2022. Evaluation of phenotype-driven gene prioritization methods for mendelian diseases. Briefings in Bioinformatics, 23(2):bbac019.

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.

Tong Zhou, Pengfei Cao, Yubo Chen, Kang Liu, Jun Zhao, Kun Niu, Weifeng Chong, and Shengping Liu. 2021. Automatic icd coding via interactive shared representation networks with self-distillation mechanism. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 5948–5957.

A Implementation Details

Span detection: for consistency with Baddour et al. (2024) and comprehensive coverage, we utilized ChatGPT-3.5 model OpenAI (2023). Corresponding prompt is described in Appendix I.

Embeddings: we used *all-MiniLM-L12-v2* Wang (2020) as the embeddings base model, from which the hyperbolic model was fine-tuned. We set *k=30* to substantially reduce the candidate space while still allowing for meaningful reranking improvements.

Candidate retrieval: FAISS Douze et al. (2024) is employed as the vector store and *Top-k* retriever for the Euclidean model. In contrast, the hyperbolic model utilizes a dedicated vector index and retrieval mechanism.

To ensure consistency in distance measurements across experiments, we normalize the hyperbolic distances in the Poincaré ball using a global normalization strategy (2):

$$\hat{d}_{\mathbb{H}}(u,v) = \frac{d_{\mathbb{H}}(u,v)}{\max_{p,q\in \text{HPO}} d_{\mathbb{H}}(p,q)} \quad (2)$$

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

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

Hybrid reranking: given an input $span$ and a set of candidates $\{C_i\}_{i=1}^{k}$, the hybrid score of a candidate C_i is computed as follows.

$$S_{\text{hybrid}}(C_i, span) = \gamma \cdot S_{\cos} - (1-\gamma)\cdot \hat{d}_{\mathbb{H}} \quad (3)$$

where $S_{\cos}$ is the cosine similarity between C_i and $span$, $\hat{d}_{\mathbb{H}}$ is their normalized hyperbolic distance, and γ is the weighting parameter between the two metrics.

B Dataset Construction

Hyperbolic Training Data

Hierarchical relationships are extracted from the HPO OWL file using DeepOnto and the ELK

reasoner. Following the methodology of He et al. (2024), we generated a dataset of triplets ⟨*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. We used random negative sampling strategy in this implementation. Given that most HPO phenotypes are associated with multiple synonyms, we augmented the dataset by including all possible synonym combinations within each 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.

Name	Definition
`WHITELIST_WORDS`	Set of medically relevant or specific terms; if any are present in a span, the span is always kept.
`BLACKLIST_PATTERNS`	Set of phrases indicating non-informative or undesirable content; if present (and not overridden by whitelist), the span is always filtered out.
`GENERIC_WORDS`	Set of common, non-specific words; used to penalize spans that are mostly generic.
`VAGUE_PATTERNS`	Set of vague or non-specific phrases; presence reduces the span's score.

Table 3: Heuristic lists used for span filtering.

Late-interaction Training Data

We fine-tuned the ColBERTv2 model Santhanam et al. (2021) on triplets of the form ⟨*span, HPO label, score*⟩, where the score represents a similarity measure. To construct a comprehensive training dataset, we first 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_sentences_spans* dataset, comprising over 200,000 clinical sentences and corresponding spans, covering the entire set of HPO terms. All prompts are provided in Appendix I for reproducibility.

After qualitative analysis of generated spans, some of them appear to be uninformative (e.g.: for the sentence "*The bladder capacity measured at 150 mL, which is below the normal range.*", the output span is "*below the normal range*" which lacks context to be relevant). We conducted a deep qualitative analysis of the spans with our clinicians in order to define filtering rules. Table 3 define the heuristics elements used in the process. The scoring procedure and the span filtering are detailed in pseudo-code in Algorithm 1 and Algorithm 2 respectively.

Algorithm 1 Compute Span Score

Require: span (string)
if span contains any **BLACKLIST_PATTERN then**
  if span contains any **WHITELIST_WORD then**
    return 1.0
  else
    return 0.0
  end if
end if
if span contains any **WHITELIST_WORD then**
  return 1.0
end if
if number of words in span ≥ 3 **then**
  score $\leftarrow$ 1.0
else
  score $\leftarrow$ 0.3
end if
generic_ratio $\leftarrow$ (number of **GENERIC_WORDS** in span) / (total words)
if generic_ratio > 0.6 **then**
  score $\leftarrow$ score -0.5
end if
if span contains any **VAGUE_PATTERN then**
  score $\leftarrow$ score -0.3
end if
return max(score, 0.0)

Algorithm 2 Filter Spans

Require: spans (list of strings)
for each span in spans **do**
  score $\leftarrow$ **Compute Span Score**(span)
  if score ≥ 0.2 **then**
    keep span
  else
    filter out span
  end if
end for

As a result, this lower-quality spans filtering yields 91,760 unique high-quality spans (with 2,167 unique spans filtered out).

Finally, we leveraged the trained hyperbolic model for scoring: positive *(span, HPO label)* pairs from the generated dataset were assigned a score

Table 4: Paired bootstrap confidence intervals (95%) for weighted Recall and weighted MRR on the ID-68 dataset. Δ denotes the mean difference between Hybrid and Euclidean retrieval ($\Delta = B - A$).

k	Metric	Euclidean (A)	Hybrid (B)	Δ [95% CI]
1	Recall	0.8663	0.8584	-0.0079 [-0.0254, $+0.0092$]
3	Recall	0.9239	0.9218	-0.0021 [-0.0113, $+0.0072$]
5	Recall	0.9340	0.9350	$+0.0009$ [-0.0078, $+0.0099$]
10	Recall	0.9452	0.9475	$+0.0024$ [-0.0045, $+0.0089$]
15	Recall	0.9544	0.9572	$+0.0028$ [-0.0016, $+0.0073$]
25	Recall	0.9621	0.9646	$+0.0025$ [-0.0023, $+0.0074$]
30	Recall	0.9643	0.9657	$+0.0014$ [-0.0023, $+0.0049$]
1	MRR	0.8663	0.8584	-0.0079 [-0.0254, $+0.0092$]
3	MRR	0.8901	0.8856	-0.0045 [-0.0165, $+0.0074$]
5	MRR	0.8919	0.8879	-0.0040 [-0.0162, $+0.0079$]
10	MRR	0.8930	0.8892	-0.0038 [-0.0155, $+0.0076$]
15	MRR	0.8931	0.8896	-0.0036 [-0.0151, $+0.0077$]
25	MRR	0.8933	0.8898	-0.0035 [-0.0150, $+0.0078$]
30	MRR	0.8933	0.8898	-0.0035 [-0.0149, $+0.0076$]

Table 5: Paired bootstrap confidence intervals (95%) for weighted Recall and weighted MRR on the CHU-50 dataset. Bold values indicate statistically significant differences (confidence interval excludes zero).

k	Metric	Euclidean (A)	Hybrid (B)	Δ [95% CI]
1	Recall	0.5966	0.6152	$\mathbf{+0.0186}$ [$\mathbf{+0.0033}$, $\mathbf{+0.0345}$]
3	Recall	0.6792	0.6899	$+0.0107$ [-0.0039, $+0.0265$]
5	Recall	0.7215	0.7234	$+0.0019$ [-0.0090, $+0.0131$]
10	Recall	0.7586	0.7592	$+0.0006$ [-0.0100, $+0.0120$]
15	Recall	0.7733	0.7782	$+0.0048$ [-0.0074, $+0.0189$]
25	Recall	0.7897	0.7984	$\mathbf{+0.0088}$ [$\mathbf{+0.0005}$, $\mathbf{+0.0189}$]
30	Recall	0.7946	0.8045	$\mathbf{+0.0099}$ [$\mathbf{+0.0021}$, $\mathbf{+0.0191}$]
1	MRR	0.5966	0.6152	$\mathbf{+0.0186}$ [$\mathbf{+0.0033}$, $\mathbf{+0.0345}$]
3	MRR	0.6279	0.6417	$\mathbf{+0.0138}$ [$\mathbf{+0.0028}$, $\mathbf{+0.0247}$]
5	MRR	0.6359	0.6474	$\mathbf{+0.0115}$ [$\mathbf{+0.0023}$, $\mathbf{+0.0207}$]
10	MRR	0.6391	0.6506	$\mathbf{+0.0115}$ [$\mathbf{+0.0022}$, $\mathbf{+0.0209}$]
15	MRR	0.6399	0.6516	$\mathbf{+0.0118}$ [$\mathbf{+0.0025}$, $\mathbf{+0.0210}$]
25	MRR	0.6404	0.6521	$\mathbf{+0.0118}$ [$\mathbf{+0.0025}$, $\mathbf{+0.0209}$]
30	MRR	0.6405	0.6522	$\mathbf{+0.0118}$ [$\mathbf{+0.0025}$, $\mathbf{+0.0208}$]

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 triplets $\langle$*span, HPO label, score*$\rangle$.

C Statistical Significance Analysis

Statistical significance was assessed using paired bootstrap resampling at the report level (2,000 resamples), with confidence intervals computed on mean differences between hybrid reranking and Euclidean retrieval.

ID-68. On the ID-68 dataset (Table 4), which contains mostly explicit phenotype mentions, hybrid reranking achieves performance comparable to Euclidean retrieval for both weighted recall and weighted MRR across all evaluated values of k. Paired bootstrap confidence intervals consistently include zero, indicating no statistically significant differences between the two approaches. This suggests that integrating hierarchical information does not degrade either retrieval coverage or ranking quality in settings where semantic similarity alone is sufficient.

CHU-50. In contrast, on the more challenging CHU-50 dataset (Table 5), which contains a higher proportion of implicit phenotype mentions, hybrid reranking yields statistically significant improvements over Euclidean retrieval. Paired bootstrap resampling confirms significant gains in weighted recall at Top-1 and at larger candidate depths, as well as consistent and statistically significant improvements in weighted MRR across all evaluated values of k (95% confidence intervals). These results demonstrate that incorporating hierarchical

structure substantially improves both retrieval and ranking of clinically relevant phenotypes in real-world clinical narratives.

D 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 6 indicates the training settings for the hyperbolic model training.

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 6: Hyperbolic Training Hyperparameters.

The training settings for the ColbertV2 fine-tuning on HPO is presented in Table 7.

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 7: ColBERTv2 Training Hyperparameters.

E Parameters Settings

For the hierarchy-aware metrics, parameters α (4) and β (5) were set empirically through random assessment by clinicians, who reviewed diverse samples to ensure clinical relevance. We verified that varying them in reasonable ranges does not change the model ranking, and all standard unweighted metrics are also reported. This ensures robustness and prevents metric-induced bias.

Regarding γ parameter (3), we conducted additional ablation studies varying the ratio between Euclidean and hyperbolic components in the hybrid model to assess their respective contributions for the ID-68 and CHU-50 dataset.

As shown in Figures 8 and 9, our findings indicate that except for very small values of γ (which consistently underperform), recall performance

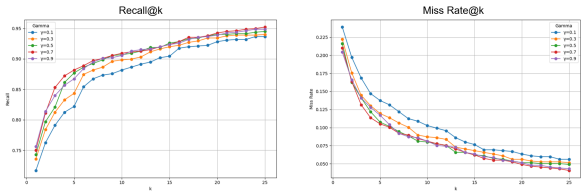

Figure 8: Recall@k and Miss Rate@k for the ID-68 dataset for different values of the weighting parameter γ, which balances Euclidean and hyperbolic distances in the hybrid reranking score.

shows no clear trends across the different γ settings. Among them, $\gamma = 0.5$ and $\gamma = 0.7$ produce the best overall results.

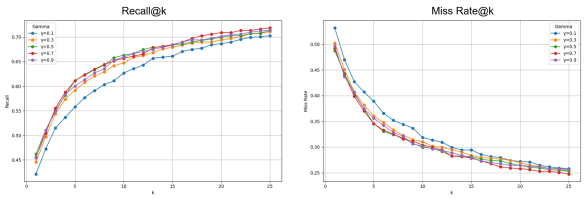

Figure 9: Recall@k and Miss Rate@k for the CHU-50 dataset for different values of the weighting parameter γ, which balances Euclidean and hyperbolic distances in the hybrid reranking score.

Interestingly, for the CHU-50 dataset (characterized by a high proportion of implicit references) the hybrid model with $\gamma = 0.5$ performs identically to, and sometimes better than, $\gamma = 0.7$. This suggests that, in such cases, the hyperbolic component is more effectively leveraged, likely due to its ability to capture hierarchical or indirect relationships.

Given these marginal differences, and to maintain simplicity and interpretability, we fix $\gamma = 0.5$ in the main experiments presented in the paper.

F Relationships Taxonomy

Figure 10 presents the type of relationships considered in the analysis of ontological relationships between the candidates and a given target phenotype (Orange). If the candidate is not in the same branch from the ontology root (*Phenotypic abnormality*), we consider that no path exist between them, as the first level of the ontology refers to very different classes of abnormalities (e.g.: Abnormality of the musculoskeletal system, Abnormality of the nervous system, Abnormality of the cardiovascular system, ...).

While *Ancestor* and *Descendant* are common types, *Cousin* is a broader notion of relationships referring to candidates that have a common ances-

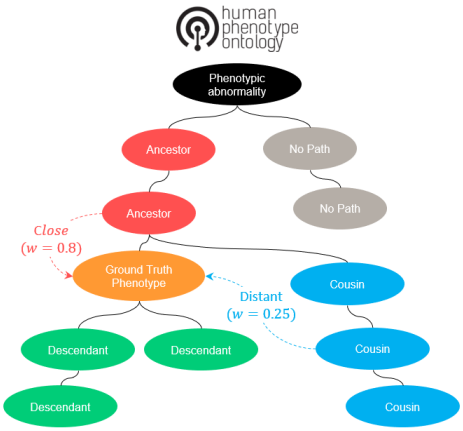

Figure 10: Relationships taxonomy. Given a target phenotype (orange), nodes are categorized as descendants (green), ancestors (red), cousins (blue), or outside the branch (grey). Relative closeness follows the relationship scores defined in Section 5.2.

tor with the target. Finally, a *Close Relationships* type refers to candidates whose relationships score (Equation 4) with the target is above a threshold of 0.7.

G Data Verification

As described in B, training data were generated using ChatGPT-4o-mini to produce sentences from HPO phenotypes with a crafted prompt 18. The generated data were manually evaluated by our clinicians.

We first randomly sampled 50 sentences ensuring coverage across different branches of the HPO ontology. Each sentence was independently reviewed by two clinicians (blind annotation). The annotation guidelines were as follows:

- Meaningfulness: The sentence should be meaningful and coherent.
- Phenotype Reference: The sentence should refer to the input phenotype, either explicitly or implicitly.
- Clinical Realism: The sentence should resemble real-world clinical report formulations.

Annotators selected among three labels: 1 for sentences fully meeting all criteria, 0 if none were met, and 0.5 for partial or ambiguous phenotype references. To assess agreement on this ordinal scale, we used quadratic-weighted Cohen's kappa, yielding a score of 0.64, indicating strong reliability despite the small, imbalanced dataset. Notably, 88% of sentences were rated as high quality. The Cohen's Kappa of 0.64 suggests a strong reliability beyond chance, especially with such small annotation set and label imbalance. The confusion matrix (Figure 11) shows strong agreement on the highest score (1), and most disagreements occur between adjacent categories (e.g., 1 vs. 0.5), suggesting that when annotators differ, they still tend to assign similar quality levels.

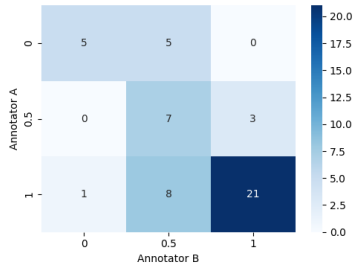

Figure 11: Annotation confusion matrix. Although the class distribution is imbalanced (majority class 1), annotators show strong agreement across classes.

We acknowledge that 50 sentences represent a small sample, but the randomized selection across diverse ontology branches and the high level of agreement observed provide a reasonably robust basis for estimating overall data quality.

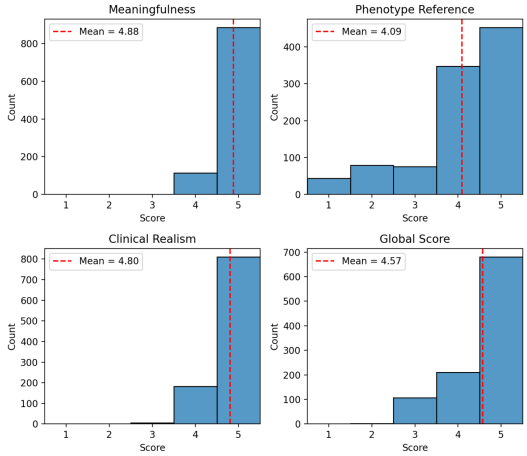

Figure 12: Distribution of LLM-as-a-judge scores across criteria. Scale are from 1 (poor quality) to 5 (excellent quality). Red lines represent the means.

To further support this assessment, we conducted an LLM-as-a-judge evaluation on a larger set of 1000 sentences across the three criteria (Figure 12). While the results should be interpreted with caution, they reinforce the observation that the vast majority of generated sentences are of high quality.

H Additional Results

Cross-Ontology Evaluation

To assess the generalizability of our workflow beyond a single ontology, we conducted additional experiments using a hyperbolic model fine-tuned on the SNOMED ontology instead of HPO. This cross-ontology evaluation tests whether our approach can leverage hierarchical structure from different sources and whether combining ontologies can improve retrieval and ranking. While both ontologies are biomedical, this experiment provides an initial step toward demonstrating the robustness and adaptability of the workflow to diverse hierarchical knowledge bases. Future work will extend this analysis to non-biomedical domains.

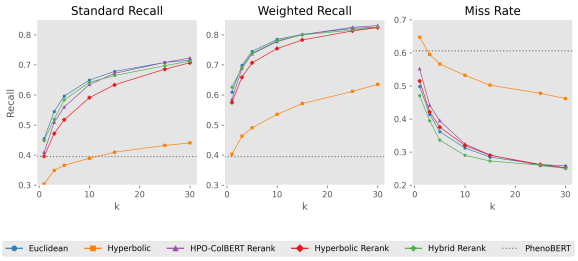

Figure 13: Recall and miss rate on ID-68 using SNOMED-trained hyperbolic embeddings. Exact-match and weighted recall are reported for raw dense retrieval (Euclidean, Hyperbolic) and reranked outputs (HPO-ColBERT, Hyperbolic, Hybrid) across *Top-k* candidates. Miss rate denotes the proportion of unretrieved ground-truth phenotypes.

As shown in Figure 13, 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 suggests that, despite the SNOMED model being less domain-specific, it may introduce complementary information that is not captured by the HPO-based hyperbolic models. The combination of semantic similarity and the broader, more general structure encoded by SNOMED embeddings could help differentiate candidates in challenging cases, leading to improved ranking performance.

Conversely, the hybrid model demonstrates reduced accuracy with SNOMED on the CHU-50 dataset. Figure 14 shows recall and miss rate for CHU-50 dataset when using SNOMED hyperbolic model. As expected, the performance is lower than with the HPO hyperbolic model. However, we can note that the miss rate is better with the hybrid approach, suggesting that the hyperbolic model's structure (even when not domain-specialized) may still be effective at ensuring broad coverage of the ontology, due to the hierarchical nature of hyperbolic embeddings.

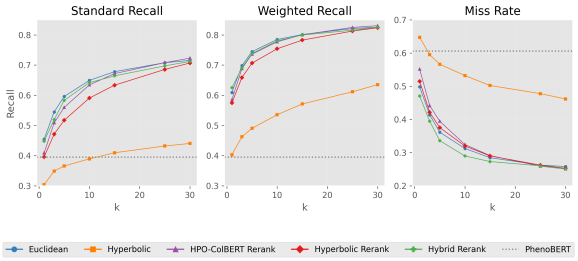

Figure 14: Recall and miss rate on CHU-50 using SNOMED-trained hyperbolic embeddings. Exact-match and weighted recall are reported for raw dense retrieval (Euclidean, Hyperbolic) and reranked outputs (HPO-ColBERT, Hyperbolic, Hybrid) across *Top-k* candidates. Miss rate denotes the proportion of unretrieved ground-truth phenotypes.

These results highlight the potential for our workflow to integrate information from multiple ontologies, supporting its applicability to a wide range of hierarchical entity linking tasks.

CHU-50 Ontology Metrics

The average number of hops between candidates and target phenotypes is shown in Figure 15. The rank scale is up to 50 so the robustness of the hyperbolic model for higher ranks is highlighted.

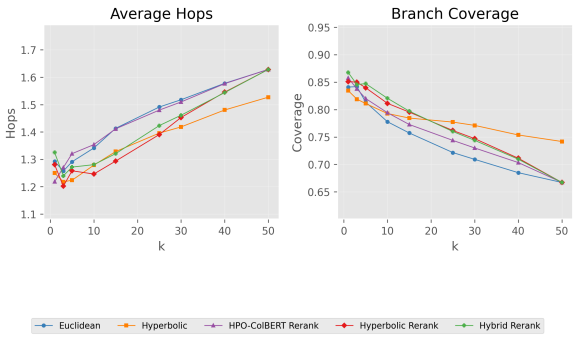

Figure 15: Ontological Structure Metrics on the CHU-50 dataset across *Top-k* candidates. Left: average number of hops between the candidates and the target phenotypes. Right: proportion of candidates within the ontology branch of the target phenotypes.

Finally, the distribution of relationships types for the CHU-50 dataset is shown in Figure 16.

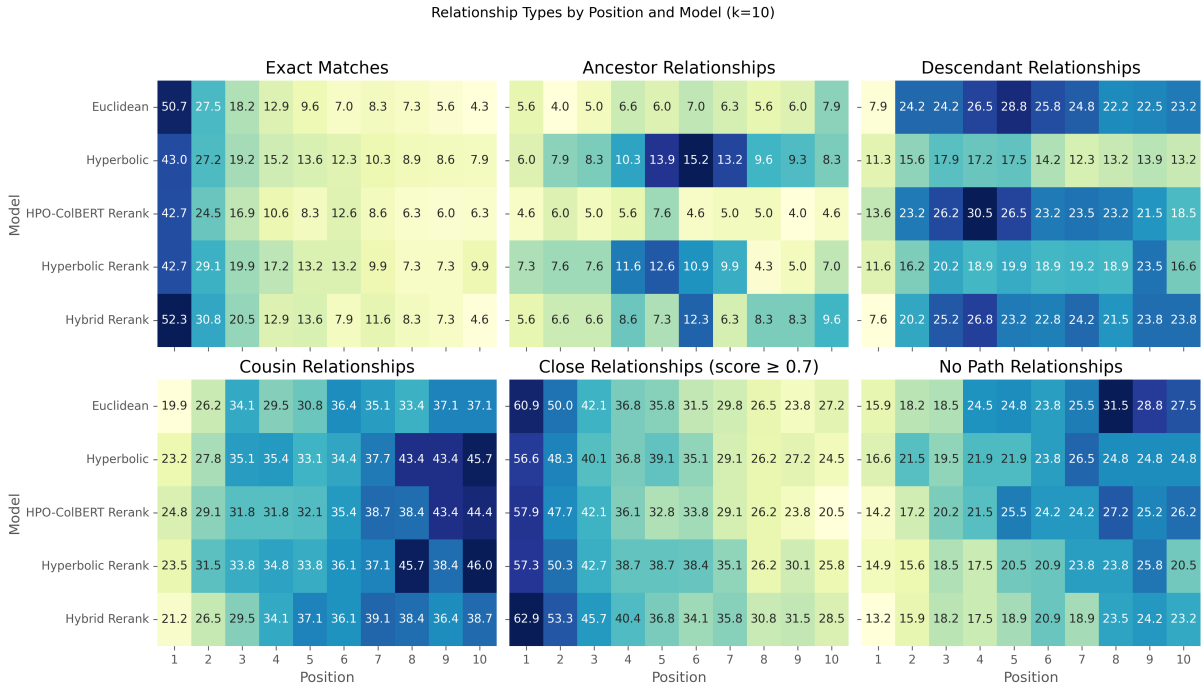

Figure 16: Distribution of relationship types by position and model for Top-10 candidates (k=10) on CHU-50 dataset. The heatmaps show percentages for six relationship categories (Appendix F) between predicted and ground truth phenotypes. For closeness relationships, only those with scores above 0.7 (Equations (4) and (5)) are considered. Rows correspond to different models, and columns represent candidate positions ranked from 1 to 10. Higher values indicate stronger presence of the relationship type at the given rank and model.

I Prompts

The prompt used for span detection step is shown in Fig. 17.

Span Detection

You are an experienced clinician with an exhaustive knowledge of human phenotypes ontology.

Given **{text}**, you must identify the spans related to possible phenotypes, either explicitly or implicitly. You should keep in the span all words related to the phenotype that should be informative (such as negation or adjective).
You may reformulate the span if needed. If you don't detect any span or if you don't know, don't try to make up an answer, just write 'None'.

Figure 17: Full prompt used for span detection (variable in blue).

Late-interaction training data are built through automatic generation of sentences from HPO. This allows to have clinically-relevant sentences for all phenotype labels and synonyms from the ontology. In addition, we crafted prompt so that implicit and measurements references are generated. The corresponding prompt is shown in Fig. 18.

As the goal was to generate spans related to phenotypes, we further query a LLM from the previously generated sentences to get the spans related to the corresponding HPO labels. The prompt used in this step is described in Fig. 19.

Finally, the prompt used for the LLM-as-a-Judge evaluation is presented in Fig. 20.

All prompts were used with the GPT-4o-mini model. The formatting preserves line breaks and spacing as passed to the model.

HPO Sentence Generation

For each HPO label, produce {sentence_count} purely observational sentences, referencing the phenotype explicitly or implicitly.

Requirements

- Use both the main HPO label and all provided synonyms for explicit references.
- At least {implicit_count} sentences must be implicit references (avoid using the label or synonyms).
- Ensure diversity in perspective or detail across sentences (as if from different medical domains and contexts).
- Ensure diversity in sentence openings.
- Use first-person ("I") or neutral ("we") style; do not include any titles (e.g., "Dr. Smith").
- At least two sentences must use passive voice.
- Vary sentence structure, wording and length (some short, some medium, some extended).
- Avoid overuse of "the patient" ($\leq$ 2 uses). Use pronouns ("he," "she," "they") or a fictitious name (first name or "Mr./Mrs." + last name) for diversity.
  - At least 1 sentence must use a fictitious name instead of "the patient" or a pronoun.
- Occasionally include measurements/tests (e.g., mg/dL, $<$ 1st percentile), up to {measures_count} total.
- No interpretive language ("suggesting," "indicative of"); only factual observations.
- No professorial explanations; just present observations.
- Return a bulleted list ("- "). Clinical shorthand is fine.

Context

- **HPO label:** {hpo_label}
- **Definition:** {definition}
- **Synonyms:** {synonyms_str}

Figure 18: Full prompt used for automatic generation of sentences from HPO (variables are shown in blue).

Span Detection from Generated Sentences

Extract the text span that best describes or indicates the phenotype "{phenotype_data['hpo_label']}" from this sentence.

Requirements

1. Capture the complete clinical observation or symptom
2. Include relevant context that helps understand the phenotype
3. Be specific to the actual medical condition
4. Be concise while maintaining clinical accuracy
5. Exclude patient names, temporal markers, or examination context
6. Focus on the actual phenotypic finding

Sentence: "{phenotype_data['sentence']}"
Extract only the relevant span without any additional commentary.

Figure 19: Full prompt used for span detection from generated sentences (variables are shown in blue). The object *phenotype_data* refers to a dataframe with the original HPO phenotype (*hpo_label*) and the corresponding generated sentence (*sentence*).

LLM-as-a-Judge Evaluation

You are an expert clinical language model evaluating sentences linked to HPO terms. Evaluate the sentence below using these criteria. Each score must be an integer from 1 to 5 (5 = excellent, 1 = poor):

1. **Meaningfulness** – Is the sentence meaningful and coherent?
2. **Phenotype Reference** – Does the sentence refer to the phenotype "**{hpo_label}**" (**{hpo_id}**), explicitly or implicitly?
3. **Clinical Realism** – Could this sentence appear in a real clinical report?

Then provide a **global score** (also 1–5) summarizing the overall quality of the sentence with respect to all three criteria.

Respond ONLY in this JSON format, filling in each `<score>` with your evaluation:

```
{
  "meaningfulness": <score>,
  "phenotype_reference": <score>,
  "clinical_realism": <score>,
  "global_score": <score>,
  "comment": "Brief explanation"
}
```

Sentence: "**{sentence}**"

Figure 20: Full prompt used for LLM-as-a-Judge evaluation.