Can Embedding Retrieval Recover Structured Compatibility?

Anonymous EMNLP submission

Abstract

We introduce OntologyBench, a controlled biomedical retrieval benchmark containing over 112k queries spanning concept grounding, relational retrieval, and compositional phenotype-based retrieval under a strict concept-level split.

Oracle ontology-aware baselines provide an upper bound and demonstrate that these tasks can be tractable under explicit enforcement of biomedical constraints. In contrast, retrieval performance degrades systematically across embedding models as tasks increasingly require jointly satisfying multiple phenotype constraints. Finetuning on ontology-derived supervision partially improves retrieval performance, while reranking and LLM-based candidate scoring provide only marginal additional gains. Our error analysis shows that embedding models frequently match phenotype features independently rather than modeling compatibility across the full phenotype set.

Together, these findings suggest that current embedding-based retrieval and downstream late-interaction or LLM-based reranking systems remain limited in their ability to recover ontology-constrained compatibility across multiple biomedical concepts. This motivates retrieval systems that can explicitly integrate representation learning with structured biomedical knowledge.

1 Introduction

Dense retrieval models enable retrieval directly over natural language by mapping semantically related queries and documents into shared embedding spaces, allowing retrieval beyond exact lexical overlap and reducing dependence on explicit ontology-grounded symbolic retrieval methods such as ontology-based semantic similarity scoring or graph traversal. This flexibility has made embedding-based retrieval attractive for biomedical and clinical applications, where concepts are often expressed through heterogeneous, incomplete, or non-canonical language.

However, many retrieval problems require satisfying multiple interdependent constraints rather than matching individual features independently. Existing benchmarks primarily evaluate retrieval under unstructured semantic similarity and provide limited insight into whether models can enforce compatibility across multiple structured inputs.

To address this gap, we introduce a retrieval evaluation framework that progressively increases compatibility requirements from concept grounding to compositional retrieval, where multiple features must be jointly satisfied. We instantiate this framework through OntologyBench, a controlled benchmark constructed from expert-curated biomedical ontologies, enabling systematic analysis of how retrieval behavior changes as constraint complexity increases.

With this framework, we identify a consistent compositional limitation: retrieval performance degrades substantially as tasks require jointly satisfying multiple structured phenotype constraints. While embedding models perform strongly on concept grounding and some pairwise ontology relations, their performance deteriorates under compositional retrieval settings, even when relevant candidates are retrieved. This pattern persists across architectures and suggests a representation-level limitation rather than a purely ranking-related deficiency. Across controlled retrieval settings, we find that embedding retrieval systems often retrieve entities matching subsets of phenotype evidence without reliably enforcing compatibility across the full phenotype combination.

This work makes three contributions: (1) We introduce a controlled retrieval framework spanning concept grounding, relational retrieval, and compositional retrieval for evaluating how embedding models handle increasingly complex biomedical compatibility constraints; (2) We show that

retrieval performance degrades systematically as tasks increasingly depend on structured and compositional compatibility, even across diverse embedding architectures and scales; and (3) Through controlled two-stage evaluation and failure analysis, we show that current embedding retrieval systems and downstream reranking remain limited in recovering ontology-constrained compositional compatibility, with failures primarily arising from independent phenotype matching rather than joint compatibility modeling.

2 Related Work

Dense Retrieval Architectures. Dense retrieval represents queries and documents as vectors in a shared embedding space, enabling efficient large-scale search (Reimers and Gurevych, 2019). Bi-encoder models support scalable retrieval via independent encoding, while late-interaction models such as ColBERT preserve token-level representations to improve fine-grained matching (Khattab and Zaharia, 2020; Chaffin and Sourty, 2025). Recent pretrained models further improve embedding quality for retrieval tasks (Zhang et al., 2025).

Embedding Benchmarks. Large-scale benchmarks such as BEIR (Thakur et al., 2021) and MTEB (Muennighoff et al., 2023) standardize evaluation across diverse retrieval and embedding tasks, primarily measuring semantic similarity in unstructured text. Biomedical retrieval datasets such as NFCorpus (Boteva et al., 2016) and TREC-COVID (Voorhees et al., 2021) similarly evaluate retrieval between clinical documents. More recent benchmarks explore retrieval requiring aggregation across multiple documents or reasoning steps, including multi-hop question answering (Yang et al., 2018) and reasoning-intensive retrieval benchmarks (Xiao et al., 2024; Su et al., 2024; Li et al., 2025). However, these benchmarks operate over unstructured text and do not isolate compositional compatibility.

Biomedical retrieval has long relied on symbolic methods that model relationships between diseases, phenotypes, and genes. Information-content–based semantic similarity methods, including Resnik similarity and phenotype set similarity approaches used in systems such as Phenomizer, leverage hierarchical biomedical structure for retrieval (Resnik, 1999; Köhler et al., 2009; Hoehndorf et al., 2015; Vasilevsky et al., 2025). Graph-based retrieval methods similarly exploit relational paths between biomedical entities. Unlike embedding-based retrieval, these approaches enforce compatibility through curated symbolic structure rather than learning it implicitly through representation similarity, making them less flexible for heterogeneous natural-language inputs such as free-text clinical documentation.

Recent work has also explored ontology-grounded biomedical embeddings designed to improve interpretability. QIME (Tang et al., 2026) constructs interpretable medical text embeddings by representing documents through ontology-grounded natural-language questions derived from biomedical concepts. While such approaches improve semantic interpretability and retrieval effectiveness, they primarily evaluate embedding quality through semantic similarity, clustering, and retrieval performance rather than through controlled evaluation of compositional compatibility. In contrast, OntologyBench focuses on whether retrieval systems can jointly satisfy multiple structured biomedical constraints under controlled retrieval settings.

OntologyBench evaluates whether embedding-based retrieval systems can recover the compatibility constraints enforced by symbolic biomedical retrieval methods. While embedding retrieval supports flexible natural-language matching, symbolic biomedical systems directly enforce compatibility through curated relational structure. OntologyBench organizes retrieval tasks from concept grounding to relational and compositional retrieval, enabling controlled analysis of how retrieval behavior changes as tasks increasingly require joint constraint satisfaction rather than independent semantic matching.

3 OntologyBench Dataset and Tasks

3.1 Knowledge Sources.

OntologyBench is designed as a controlled evaluation framework that isolates how embedding models behave under increasing levels of structured and compositional constraints. By organizing tasks from concept grounding to compositional retrieval and enforcing concept-level generalization, OntologyBench evaluates how retrieval models handle increasingly complex compatibility constraints.

We construct a structured retrieval space grounded in expert-curated ontologies linking clinical findings, diseases, and genetic causes. Disease concepts are derived from the Monarch Disease On-

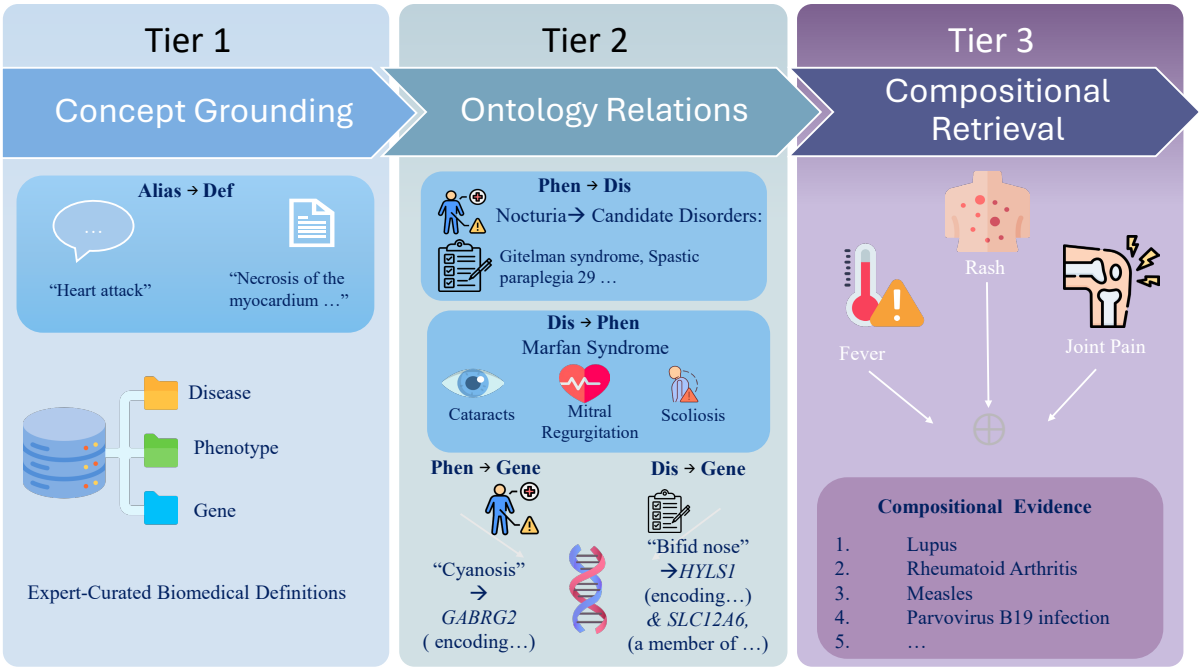

Figure 1: OntologyBench overview illustrating the retrieval tiers from grounding to compositional retrieval.

tology (MONDO) (Vasilevsky et al., 2025), which harmonizes disease definitions across biomedical vocabularies such as International Classification of Diseases (ICD), Systematized Nomenclature of Medicine—Clinical Terms (SNOMED CT), and Online Mendelian Inheritance in Man (OMIM). OntologyBench includes 6,696 canonical MONDO disease concepts, each associated with one or more textual aliases.

Phenotype concepts are drawn from the Human Phenotype Ontology (HPO) (Gargano et al., 2024), a controlled vocabulary describing observable clinical features (phenotypes). The dataset contains 8,467 phenotype concepts, each with associated textual synonyms used as query inputs.

Gene entities are constructed from HUGO Gene Nomenclature Committee(HGNC) gene names (Seal et al., 2026) combined with National Center for Biotechnology Information (NCBI) Gene metadata (O'Leary et al., 2024), resulting in 2,345 gene entries.

Together, these ontologies provide a structured biomedical knowledge space linking phenotypes, diseases, and genes, enabling evaluation of retrieval systems under structured biomedical constraints.

3.2 Task Formulation

OntologyBench is organized into three tiers spanning concept grounding, ontology relations, and compositional retrieval (Figure 1). This design enables controlled evaluation of how retrieval behavior changes as tasks require increasingly complex compatibility constraints. Examples of each tier can be found in Appendix A.

Tier 1: Concept Grounding Tier 1 evaluates whether models can map surface forms (aliases) of genes (Gene), phenotypes (Phen), and disorders (Dis) to their canonical ontology definitions (Def) through single-hop retrieval. This tier includes three grounding tasks: **Gene→Def**, **Phen→Def**, and **Dis→Def**.

Tier 2: Relational Retrieval. This tier evaluates whether models capture pairwise biomedical relationships between diseases, phenotypes, and genes.

Disease ↔ Phenotype (Dis↔Phen) These tasks evaluate retrieval over curated disease–phenotype associations in both directions. Disease-to-phenotype retrieval maps disorders to characteristic features, whereas phenotype-to-disease retrieval is more ambiguous because many phenotypes are shared across diseases.

Phenotype→Gene (Phen↔Gene). Given a phenotype query, the model retrieves genes associated with diseases exhibiting that phenotype.

Disease→Gene (Dis↔Gene). Given a disease query, the model retrieves genes known to be molecularly associated with that disorder.

Task	Train #Queries	Train AvgPos	Test #Queries	Test AvgPos
Tier 1 (Concept Grounding)				
Gene→Def	1,914	1.00	431	1.00
Phen→Def	17,763	1.00	4,581	1.00
Dis→Def	39,644	1.00	9,205	1.00
Dis→Phen	6,647	19.81	6,291	5.63
Phen→Dis	7,988	16.62	5,229	6.57
Tier 2 (Relational Retrieval)				
Phen→Gene	6,960	1.00	1,507	1.00
Dis→Gene	6,623	17.58	6,504	5.31
Tier 3 (Compositional Retrieval)				
PhenTriplet→Dis	5,642	4.39	2,537	2.26

Table 1: OntologyBench dataset statistics across tiers. AvgPos denotes the average number of positive documents per query.

Tier 3: PhenTriplet→Dis (compositional retrieval). In clinical settings, diagnosis is often based on incomplete and partially observed phenotypes. Tier 3 models this under a controlled phenotype-based retrieval setting, evaluating compositional retrieval from limited evidence. Each query consists of three phenotypes drawn from a single disease's phenotype set, ensuring internal consistency. Some phenotype subsets may include non-specific features; this is intentional, as such cases introduce ambiguity and test whether models can correctly aggregate weak signals rather than rely on a single highly discriminative feature.

The choice of three phenotypes reflects a trade-off between ambiguity and determinism: single phenotypes correspond to many candidate diseases, while larger sets (≥ 4 phenotypes) sharply reduce the candidate space, making retrieval increasingly deterministic. As shown in Appendix B, phenotype triplets yield a non-trivial regime (~ 5 candidate diseases on average), requiring models to aggregate multiple constraints rather than rely on single-feature matching.

While this construction does not model full clinical complexity, it isolates compositional aggregation under controlled conditions, enabling evaluation of whether models can integrate multiple phenotype signals under ambiguity.

3.3 Concept-Level Partition

The unified multi-tier corpus contains 471,854 training pairs and 125,744 evaluation pairs, constructed under an 80/20 concept-level partition. The split is performed over ontology concepts rather than individual query–document pairs, ensuring that evaluation concepts are absent from training and enforcing strict unseen-concept generalization, thereby preventing memorization of ontology entities or relations and requiring models to generalize structured and compositional relationships to unseen concepts.

3.4 Dataset Statistics

Table 1 summarizes dataset statistics across tiers. (Additional dataset statistics analysis can be found in B). The benchmark contains structured retrieval queries covering concept grounding, biomedical associations, and compositional retrieval.

Tier 1 grounding tasks are deterministic (AvgPos = 1), while Tier 2 ontology association tasks exhibit substantial ambiguity due to many-to-many biomedical relationships (e.g., Dis→Phen and Phen→Dis average 19.8 and 16.6 positives per query). In contrast, Tier 3 compositional queries correspond to a smaller candidate set (4.39 diseases on average), reflecting the narrowing effect of combining multiple phenotype constraints. Sequence length statistics are reported in Appendix C.

This progression establishes a controlled transition from deterministic to compositional retrieval, enabling analysis of how model behavior changes as aggregation requirements increase.

4 Method

4.1 Supervision Settings

To study how ontology supervision shapes embedding representations, we compare two training regimes under the same concept-level train–test partition.

Concept Grounding Supervision. Models are trained only on alias–definition grounding pairs that map surface biomedical expressions to canonical concepts. No ontology relations or compositional signals are provided during training. Performance on higher-tier tasks, therefore, measures whether relational structure can emerge from concept grounding alone.

Unified Ontology Supervision. Models are trained jointly on all ontology-derived tasks, including grounding, entity relations (e.g., phenotype–disease, disease–gene), and compositional phenotype queries. Because evaluation concepts are entirely unseen during training, models must generalize relational and compositional knowledge to new biomedical entities.

4.2 Two-Stage Retrieval Evaluation

Performance degradation under compositional retrieval may arise from either missing candidates during retrieval or incorrect downstream ranking among retrieved candidates. To disentangle these factors, we introduce a controlled two-stage retrieval evaluation framework.

First, an initial embedding retriever generates the top-50 candidate set from the full corpus (Recall@50 = 0.793). Second, we evaluate whether downstream reranking can recover the correct diagnosis from this fixed high-recall candidate pool. Late-interaction reranking is implemented using GTE-ModernColBERT with ColBERT-style MaxSim scoring, while LLM-based candidate scoring evaluates phenotype–disease compatibility directly over the retrieved candidates (Appendix F).

Together, these results suggest that current embedding retrieval and reranking approaches remain insufficient for reliably recovering ontology-constrained biomedical compatibility from unstructured text alone.

4.3 Experimental Setup

Model Setup. We evaluate lexical, dense retrieval, reranking, and generative models. As a lexical baseline, we use BM25. Dense retrieval models include general-purpose embedding models such as all-MiniLM-L6-v2 (Reimers and Gurevych, 2019), Qwen3 embeddings (0.6B and 4B) (Zhang et al., 2025), and OpenAI `text-embedding-3-large`[1], as well as biomedical embedding models including BioLORD-0.1B (Remy et al., 2024), MedEmbed-0.1B (Balachandran, 2024), and SapBERT-0.1B (Liu et al., 2021).

For late-interaction reranking, we evaluate GTE-ModernColBERT (Chaffin and Sourty, 2025) using ColBERT-style MaxSim scoring (Khattab and Zaharia, 2020). For LLM-based candidate scoring reranking, we use Qwen3-4B-Instruct-2507, Qwen3-4B-Thinking-2507 (Team, 2025), Qwen3.6-27B (Qwen Team, 2026), Gemma4-31B (Google DeepMind, 2026), and OpenAI GPT5.4-mini[2].

Oracle Ontology-Aware Baselines. To evaluate whether embedding-based retrieval can approximate the compatibility constraints enforced explicitly by symbolic ontology-aware systems, we include oracle structure-aware baselines operating directly over curated biomedical structures.

For phenotype-related tasks (Dis→Phen, Phen→Dis, PhenTriplet→Dis), we use a Phenomizer-style semantic similarity method based on information content with set-level aggregation over HPO phenotypes (Köhler et al., 2009).

For gene-related tasks (Phen→Gene, Dis→Gene), where ontology-based semantic similarity is not applicable because genes are not represented within the HPO hierarchy, we use a graph traversal baseline based on disease–gene connectivity, aggregating evidence along phenotype–disease–gene paths. For Dis→Gene retrieval, ranking reduces to retrieving genes directly associated with the query disease under the ontology graph.

These oracle baselines operate under idealized structure-aware settings with direct access to curated biomedical relationships, and are therefore intended as diagnostic reference baselines rather than directly comparable text-only retrieval systems.

Training Objective. Fine-tuned retrieval models are trained using Cached Multiple Negatives Ranking Loss (CachedMNRL), a contrastive objective with in-batch negatives (Lindgren et al., 2021):

$$\mathcal{L}(q, d^+) = -\log \frac{\exp(\mathrm{sim}(q, d^+))}{\sum_{d \in \mathcal{B}} \exp(\mathrm{sim}(q, d))}, \quad (1)$$

where $\mathcal{B}$ denotes the training batch. Similarity is computed using cosine similarity for bi-encoder models, while late-interaction models use MaxSim aggregation.

Training Setup. To ensure fair comparison across supervision regimes, training budgets and optimization procedures are kept identical for all models. More training details are provided in Appendix D.

4.4 Evaluation Metrics

Retrieval performance is evaluated using nDCG, MRR, and Recall@k ($k \in \{1, 5, 10\}$). We report nDCG@10 in the main paper and provide additional metrics and details in Appendix E. Relevance is binary, and metrics are computed over the full candidate set for each task and averaged across queries.

[1] https://platform.openai.com/docs/guides/embeddings

[2] https://developers.openai.com/api/docs/guides/reasoning

Task	Oracle Baseline	nDCG@10
1B-R (Dis→Phen)	Phenomizer-style	0.716
1C-R (Phen→Dis)	Phenomizer-style	0.831
2A-R (Phen→Gene)	Graph Traversal	0.840
2B-R (Dis→Gene)	Graph Traversal	0.933
3A (PhenTriplet→Dis)	Phenomizer-style	0.778

Table 2: Performance of oracle ontology-aware baselines evaluated using nDCG@10.

While we use standard retrieval metrics, ontology constraints are enforced through the task construction: queries and candidate sets are derived from curated biomedical relations and compositional phenotype structures. As a result, these metrics evaluate whether models retrieve and rank entities consistent with structured biomedical relationships rather than purely semantic similarity.

5 Results

5.1 Oracle Structure-Aware Retrieval Reveals a Large Gap in Compositional Compatibility

Oracle baselines demonstrate that OntologyBench retrieval tasks are highly solvable when explicit biomedical structure is directly accessible. As shown in Table 2, ontology-aware methods operating over curated hierarchical and relational structure achieve strong retrieval performance across phenotype-based and compositional tasks (e.g., PhenTriplet→Dis nDCG@10 = 0.778).

These oracle systems operate under fundamentally different assumptions from embedding-based retrieval models. Rather than inferring compatibility from textual representations, they directly exploit explicit ontology hierarchy structure or curated disease–gene connectivity. The oracle baselines are therefore not intended as direct competitors, but instead serve as diagnostic references establishing the level of retrieval performance achievable when structured biomedical compatibility is explicitly available.

The oracle baselines reveal a large gap between explicit ontology-guided information lookup and modern embedding retrieval systems. While ontology-aware methods directly exploit curated biomedical relationships and hierarchical structure, embedding models must recover such compatibility implicitly through representation learning from unstructured text and retrieval supervision alone. The observed performance gap suggests that current embedding training paradigms remain insufficient for reliably capturing structured-constrained compatibility.

5.2 Performance Degrades as Retrieval Requires Stronger Compatibility Constraints

As shown in Table 3, embedding model performance declines as retrieval tasks require increasingly complex compatibility constraints. Performance is highest on Tier-1 grounding tasks compared with higher-complexity retrieval tasks, but remains imperfect across models and task types. While some tasks (e.g., Phen→Def) achieve strong performance, others remain substantially lower, indicating that even single-entity grounding is not uniformly solved.

The drop becomes more pronounced under relational and compositional queries, where multiple constraints must be satisfied jointly. These results reveal a clear pattern: retrieval performance degrades as tasks require modeling more complex structured relationships.

5.3 Domain-Specific Models Improve Relational Retrieval but Do Not Resolve Compositional Challenges

Across retrieval complexity levels, domain-specialized embedding models consistently outperform larger general-purpose embeddings on relational and compositional retrieval tasks. In particular, BioLORD-0.1B surpasses both Qwen3-Embed-0.6B and Qwen3-Embed-4B on compositional phenotype-based disease retrieval despite its substantially smaller size, while Qwen3-Embed-4B performs worse than the lexical BM25 baseline in the compositional setting.

These results suggest that model scaling alone is insufficient for retrieval requiring structured biomedical compatibility. Although ontology-aligned biomedical supervision improves relational retrieval, substantial compositional errors remain when multiple phenotype constraints must be jointly satisfied.

5.4 Task-Aligned Supervision Outweighs Model Scale in Structured Biomedical Retrieval

Unified multi-tier supervision substantially improves performance on relational and compositional retrieval tasks. Across tasks, fine-tuned models consistently outperform the strongest pretrained baselines, with gains of up to +0.255 nDCG@10 on Dis→Gene retrieval and +0.144 on compositional

Model	nDCG@10							
	Tier 1			Tier 2				Tier 3
	Gene→Def	Phen→Def	Dis→Def	Dis→Phen	Phen→Dis	Phen→Gene	Dis→Gene	PhenTriplet→Dis
BM25	0.087	0.381	0.275	0.155	0.178	0.071	0.140	0.138
MiniLM-L6-v2	0.263	0.629	0.314	0.227	0.249	0.109	0.146	0.148
MedEmbed-0.1B	0.440	0.738	0.454	0.269	0.268	0.134	0.171	0.167
SapBERT-0.1B	0.670	0.818	0.476	0.194	0.257	0.113	0.168	0.160
BioLORD-0.1B	0.290	0.807	0.426	0.201	0.285	0.105	0.143	0.182
BioLORD-0.1B†	0.513	**0.869**	**0.562**	0.245	0.276	0.091	0.123	0.204
BioLORD-0.1B‡	0.455	0.668	0.532	0.333	0.250	**0.191**	0.213	0.283
Qwen3-Embed-0.6B	0.194	0.449	0.174	0.343	0.257	0.058	0.264	0.098
Qwen3-Embed-0.6B†	0.459	0.796	0.470	0.392	0.250	0.046	0.180	0.138
Qwen3-Embed-0.6B‡	0.331	0.498	0.434	**0.586**	**0.414**	0.183	**0.546**	**0.344**
Qwen3-Embed-4B	0.354	0.457	0.203	0.433	0.305	0.071	0.291	0.064
OpenAI (text-embedding-3-large)	**0.747**	0.860	0.516	0.248	0.303	0.136	0.170	0.200

Table 3: Retrieval performance on OntologyBench, evaluated using nDCG@10 across eight retrieval tasks spanning three retrieval tiers from grounding to compositional retrieval. Fine-tuned models are highlighted in blue, where † indicates models fine-tuned on the Tier 1 training set only, and ‡ indicates models fine-tuned jointly on the training sets from all three tiers.

PhenTriplet→Dis. These improvements indicate that ontology-aligned supervision enables models to better capture structured relationships between biomedical concepts, particularly in settings requiring integration of multiple constraints. Notably, smaller fine-tuned models outperform larger pre-trained embeddings, suggesting that task-aligned supervision is more critical than model scale for ontology-constrained retrieval. Additional evaluation results can be found in Appendix G.

5.5 Late Interaction and Generative Reranking Do Not Resolve Compositional Retrieval Failures

Late-interaction rerankers substantially underperform the first-stage retriever on Tier-3 compositional retrieval tasks, and reranker fine-tuning further degrades performance (Table 4). LLM-based candidate scoring performs better than token-level reranking and slightly exceeds the retriever itself, with the strongest model (GPT5.4-mini) improving nDCG@10 from 0.344 to 0.362. However, these gains remain modest. Together, the reranking results suggest that the primary limitation lies in the retrieval representations themselves—specifically, the inability to reliably model joint compatibility across multiple phenotypes—rather than insufficient downstream ranking capacity alone.

5.6 Failure Modes in Compositional Retrieval

We perform a rule-based error analysis (Appendix H) on Tier-3 queries where embedding retrieval fails, but the ontology-aware baseline succeeds. Failures cluster into four modes: (1) **aggregation failure** (58.5%), where the model retrieves

Model	$k = 10$		
	MRR	**Recall**	**nDCG**
Retriever			
Qwen3-Embed-0.6B‡	0.286	0.547	0.344
Token-Interaction Rerankers			
ModernColBERT	0.168	0.384	0.217
ModernColBERT†	0.168	0.386	0.218
ModernColBERT‡	0.135	0.384	0.192
LLM Candidate Scoring			
Qwen3-4B-Instruct	0.261	0.513	0.318
Qwen3-4B-Think	0.279	0.546	0.339
Qwen3.6-27B	0.297	0.572	0.361
Gemma4-31B	0.289	0.570	0.355
OpenAI (GPT5.4-mini)	**0.301**	**0.574**	**0.362**

Table 4: Two-stage evaluation on Tier-3 compositional retrieval tasks isolating downstream ranking from candidate retrieval. Despite fixing the candidate set, reranking and LLM-based scoring provide only marginal improvement over the retriever, suggesting that compositional failures primarily originate from the underlying retrieval representations.

diseases matching only a subset of the query phenotypes rather than the full phenotype combination; (2) **generic phenotype bias** (25.0%), where retrieval is driven by broad or non-specific phenotypes; (3) **related-disease confusion** (11.8%), where the predicted disease is phenotypically similar to the ground-truth disease; and (4) **semantic drift** (4.7%), where retrieved diseases show low phenotype overlap and low ontology similarity. Aggregation failures dominate, indicating that embedding models often match phenotypes independently rather than enforcing joint compatibility across the full phenotype set.

Ontology-aligned supervision substantially reduces catastrophic retrieval failures where the ground-truth disorder is absent from retrieval

(70.6%→29.4%), while improving agreement with ontology-aware retrieval rankings, although compositional ranking errors remain common. (Appendix H.1).

Overall, embedding models retrieve partially relevant candidates but fail to satisfy multi-phenotype constraints, even when correct answers are present. This limitation is consistent with lower performance on relation types requiring multi-hop reasoning (e.g., phenotype→gene), indicating that compositional structure remains underrepresented in current representations.

6 Discussion

Retrieval performance degrades consistently as tasks increasingly require jointly satisfying structured biomedical constraints, with the largest drop observed in multi-phenotype compositional retrieval. This pattern persists across embedding architectures, suggesting a general representational limitation rather than a model-specific failure. Although ontology-aligned supervision partially improves relational and compositional retrieval, substantial errors remain, indicating that current similarity-based retrieval paradigms only partially capture structured biomedical compatibility.

Oracle structure-aware baselines further demonstrate that these tasks are highly tractable when explicit biomedical structure is directly accessible. In contrast, embedding retrieval systems must recover such compatibility implicitly through representation learning from unstructured text and retrieval supervision alone. The resulting performance gap suggests that current retrieval paradigms remain insufficient for reliably modeling ontology-constrained biomedical compatibility.

Our two-stage retrieval evaluations further localize this limitation to the learned representations themselves. Even when the correct diagnosis is retrieved within a high-recall candidate pool, neither late-interaction reranking nor LLM-based candidate scoring substantially improves performance, suggesting that increased downstream reasoning capacity alone does not resolve the underlying compositional representation failure.

Failure analysis further supports this interpretation. Aggregation failure accounts for 58.5% of compositional retrieval errors, indicating that models frequently match phenotype features independently rather than enforcing compatibility across the full phenotype set.

7 Conclusion

We introduced OntologyBench, a controlled benchmark for evaluating biomedical retrieval under grounding, relational, and compositional constraints. Across embedding models, retrieval performance degrades substantially as tasks increasingly require jointly satisfying multiple phenotype constraints, while reranking provides only limited improvement. In contrast, oracle structure-aware retrieval methods remain highly effective under explicit biomedical constraint enforcement.

Together, these findings suggest that current embedding retrieval and downstream LLM-based reranking paradigms remain insufficient for reliably modeling ontology-constrained biomedical compatibility from unstructured text alone. We hope OntologyBench facilitates future research on retrieval systems that more effectively integrate representation learning with structured biomedical knowledge.

8 Data and Code Availability

Code, models, and benchmark data will be publicly released upon acceptance.

Limitations

OntologyBench evaluates retrieval under controlled biomedical compatibility constraints and does not capture broader aspects of real-world diagnosis, such as temporal progression, causal mechanisms, incomplete observations, or patient-specific context.

Although concept-level partitioning prevents direct memorization of ontology entities and relations, textual definitions may still share recurring lexical patterns that provide weak indirect cues for relational structure. However, retrieval performance degrades consistently as tasks increasingly require integrating multiple compatibility constraints (Table 3), particularly under compositional retrieval settings that require aggregating multiple partially informative phenotypes. This suggests that lexical overlap alone is insufficient to explain higher-level retrieval performance, though controlling more explicitly for lexical similarity remains important future work.

Our experiments evaluate a representative range of retrieval architectures and model scales, but are not intended as an exhaustive leaderboard-style comparison of all emerging foundation systems.

The primary goal of OntologyBench is instead to characterize retrieval behavior as tasks increasingly require structured and compositional compatibility. Future work may investigate whether larger-scale or ontology-aware architectures mitigate the limitations identified here.

OntologyBench evaluates retrieval behavior under structured biomedical constraints rather than directly measuring preservation of ontology geometry within embedding space. Standard retrieval metrics such as nDCG, MRR, and Recall@k therefore do not explicitly quantify hierarchical distance preservation or geometric alignment with ontology topology. Consequently, the observed degradation under compositional and relational retrieval settings should be interpreted as evidence of limitations in compositional retrieval rather than a definitive measure of ontology structure preservation within learned representations.

Although ontology-aligned supervision improves relational and compositional retrieval, the current study does not fully isolate which components of the supervision signal contribute most strongly to these gains. Improvements may arise from a combination of structured relational supervision, task-specific retrieval supervision, and domain adaptation effects. Future work could further disentangle these factors through more controlled supervision ablations.

Finally, OntologyBench focuses on representation-level retrieval. While practical clinical systems may incorporate reranking, generative reasoning, and expert validation, our results suggest that substantial limitations already emerge at the retrieval stage. The benchmark should therefore be interpreted as a diagnostic framework for analyzing retrieval behavior under structured biomedical constraints rather than a direct proxy for end-to-end clinical performance.

Ethical Considerations

OntologyBench is constructed entirely from publicly available biomedical ontologies released under open licenses, including the HPO and MONDO (CC-BY 4.0) and HGNC and NCBI resources (public domain). These sources provide structured definitions of concepts and relations rather than patient-level data. OntologyBench contains no personally identifiable information or sensitive health records.

Statement of AI Use

OpenAI (GPT-4) was used for language polishing. All research content, experimental design, analysis, and conclusions were written and verified by the authors.

References

Abhinand Balachandran. 2024. Medembed: Medical-focused embedding models.

Vera Boteva, Demian Gholipour, Artem Sokolov, and Stefan Riezler. 2016. A full-text learning to rank dataset for medical information retrieval. In *European Conference on Information Retrieval*, pages 716–722. Springer.

Antoine Chaffin and Raphaël Sourty. 2025. Pylate: Flexible training and retrieval for late interaction models. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pages 6334–6339.

Michael A Gargano, Nicolas Matentzoglu, Ben Coleman, Eunice B Addo-Lartey, Anna V Anagnostopoulos, Joel Anderton, Paul Avillach, Anita M Bagley, Eduard Bakštein, James P Balhoff, et al. 2024. The human phenotype ontology in 2024: phenotypes around the world. *Nucleic acids research*, 52(D1):D1333–D1346.

Google DeepMind. 2026. gemma-4-31b. https://huggingface.co/google/gemma-4-31B. Hugging Face model repository.

Robert Hoehndorf, Paul N Schofield, and Georgios V Gkoutos. 2015. The role of ontologies in biological and biomedical research: a functional perspective. *Briefings in bioinformatics*, 16(6):1069–1080.

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 39–48.

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

Lei Li, Xiao Zhou, and Zheng Liu. 2025. R2med: A benchmark for reasoning-driven medical retrieval. *arXiv preprint arXiv:2505.14558*.

Erik Lindgren, Sashank Reddi, Ruiqi Guo, and Sanjiv Kumar. 2021. Efficient training of retrieval models using negative cache. *Advances in Neural Information Processing Systems*, 34:4134–4146.

Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. 2021. Self-alignment pretraining for biomedical entity representations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4228–4238, Online. Association for Computational Linguistics.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2023. Mteb: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037.

Nuala A O'Leary, Eric Cox, J Bradley Holmes, W Ray Anderson, Robert Falk, Vichet Hem, Mirian TN Tsuchiya, Gregory D Schuler, Xuan Zhang, John Torcivia, et al. 2024. Exploring and retrieving sequence and metadata for species across the tree of life with ncbi datasets. *Scientific data*, 11(1):732.

Qwen Team. 2026. Qwen3.6-27B: Flagship-level coding in a 27B dense model.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

François Remy, Kris Demuynck, and Thomas Demeester. 2024. BioLORD-2023: semantic textual representations fusing large language models and clinical knowledge graph insights. *Journal of the American Medical Informatics Association*, page ocae029.

Philip Resnik. 1999. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of artificial intelligence research*, 11:95–130.

Ruth L Seal, Bryony Braschi, Kristian Gray, James McClay, Susan Tweedie, and Elspeth A Bruford. 2026. Genenames. org: the hgnc and pgnc resources in 2026. *Nucleic Acids Research*, 54(D1):D1098–D1107.

Hongjin Su, Howard Yen, Mengzhou Xia, Weijia Shi, Niklas Muennighoff, Han-yu Wang, Haisu Liu, Quan Shi, Zachary S Siegel, Michael Tang, et al. 2024. Bright: A realistic and challenging benchmark for reasoning-intensive retrieval. *arXiv preprint arXiv:2407.12883*.

Yixuan Tang, Zhenghong Lin, Yandong Sun, Wynne Hsu, Mong Li Lee, and Anthony KH Tung. 2026. Qime: Constructing interpretable medical text embeddings via ontology-grounded questions. *arXiv preprint arXiv:2603.01690*.

Qwen Team. 2025. Qwen3 technical report.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models. *arXiv preprint arXiv:2104.08663*.

Nicole A Vasilevsky, Sabrina Toro, Nicolas Matentzoglu, Joseph E Flack, Kathleen R Mullen, Harshad Hegde, Sarah Gehrke, Patricia L Whetzel, Yousif Shwetar, Nomi L Harris, et al. 2025. Mondo: Integrating disease terminology across communities. *Genetics*, page iyaf215.

Ellen Voorhees, Tasmeer Alam, Steven Bedrick, Dina Demner-Fushman, William R Hersh, Kyle Lo, Kirk Roberts, Ian Soboroff, and Lucy Lu Wang. 2021. Trec-covid: constructing a pandemic information retrieval test collection. In *ACM SIGIR Forum*, volume 54, pages 1–12. ACM New York, NY, USA.

Chenghao Xiao, G Thomas Hudson, and Noura Al Moubayed. 2024. Rar-b: Reasoning as retrieval benchmark. *arXiv preprint arXiv:2404.06347*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 2369–2380.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, et al. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.

A Dataset Examples

This appendix shows representative examples from the three tiers of OntologyBench tasks.

A.1 Tier 1: Concept Grounding

Gene→Document Retrieval

doc_id: GENE:2011
doc: Protein-coding gene located on chromosome 11q13.1 belonging to the microtubule affinity regulating kinase family. The encoded kinase regulates epithelial and neuronal cell polarity and microtubule stability.

Phen→Definition Retrieval

doc_id: HP:0006349
doc: Congenital absence of one or more permanent teeth, including hypodontia, oligodontia, or complete tooth absence.

MONDO→Definition Retrieval

doc_id: :MONDO:0014543
doc: Congenital myasthenic syndrome with glycosylation defect caused by mutations in the ALG2 gene.

A.2 Tier 2: Ontology-Level Relations

Disease→Phenotype

doc_id: HP:0002154
doc: Elevated concentration of glycine in the blood.

Phenotype→Disease

doc_id: MONDO:0009479
doc: Johanson-Blizzard syndrome characterized by exocrine pancreatic insufficiency, nasal alae hypoplasia, hearing loss, growth retardation, and intellectual disability.

Phenotype→Gene

doc_id: GENE:80185
doc: TTI2 encodes a regulator of the DNA damage response and a component of the Triple T complex involved in cellular resistance to DNA damage stresses.

Disease→Gene

doc_id: GENE:889
doc: KRIT1 encodes a protein involved in beta1-integrin-mediated cell proliferation and endothelial junction integrity. Mutations cause cerebral cavernous malformations.

A.3 Tier 3: Compositional Phenotype-Based Disease Retrieval

Multi-Phenotype→Disease

doc_id: MONDO:0016394
doc: Sporadic infantile bilateral striatal necrosis characterized by degeneration of the caudate nucleus, putamen, and globus pallidus, leading to developmental regression and movement disorders.

Multi-Phenotype→Disease

doc_id: MONDO:0011614
doc: HMG-CoA synthase-2 deficiency, a rare disorder of ketone body metabolism presenting with vomiting, lethargy, hepatomegaly, and non-ketotic hypoglycemia.

Multi-Phenotype→Disease

doc_id: MONDO:0020502
doc: Yellow fever, a zoonotic viral disease that may progress from fever and systemic symptoms to hemorrhagic fever and multi-organ failure.

B Additional Dataset Analysis

Tier-1 Statistics Disease identity grounding (MONDO) contains 6,611 canonical concepts associated with 48,854 alias queries (mean 7.39 aliases per concept; median 6; maximum 50).

Alias polysemy is limited: 1.0% of surface forms map to more than one concept. Gene identity grounding contains 2,345 gene concepts, each associated with a single canonical symbol. Phenotype grounding includes 8,253 phenotype concepts with extensive synonym coverage.

These statistics indicate that Tier-1 primarily evaluates lexical normalization under moderate alias diversity rather than heavy ambiguity.

Phenotype Cardinality and Hypothesis Space Reduction We fix Tier-3 phenotype cardinality to three to study the minimal non-trivial regime of compositional retrieval. Ontology association statistics indicate:

- Single phenotype: mean 328.9 compatible diseases
- Phenotype pairs: mean 26.7 compatible diseases
- Phenotype triplets: mean 5.0 compatible diseases
- Phenotype quadruplets: mean 1.95
- Phenotype quintuplets: mean 1.31

Triplets provide substantial hypothesis-space reduction while avoiding near-deterministic collapse observed at higher cardinalities.

C Experimental Details

Model	Mean	P95	P99	Max
Qwen3-Embed-0.6B	33.40	112	161	1053
BioLORD-0.1B	48.39	128	205	926
ColBERT	25.08	88	149	845

Table 5: Token length statistics of training instances under different tokenizers (no truncation applied). We report the mean length, 95th percentile (P95), 99th percentile (P99), and maximum sequence length.

Sequence Length Distribution and Truncation All models were trained with a per-device batch size of 128 and gradient accumulation of 1. Training was conducted for a fixed schedule of 2 epochs across all architectures to ensure equal training exposure.

Bi-encoder models used a maximum sequence length of 256 tokens to control truncation under the heavy-tailed definition length distribution. The ModernColBERT reranker retained its native sequence configuration (query length of 48 and document length of 300) to preserve its architecture-specific interaction design. All models were trained on the same single 40GB GPU hardware configuration.

D Optimization and Training Setup

All models were fine-tuned for two epochs using the AdamW optimizer with an effective batch size of 256. Bi-encoder models (BioLORD-0.1B and Qwen3-Embedding-0.6B) were trained using Cached Multiple Negatives Ranking Loss (CachedMNRL), which enables large in-batch negative pools by caching intermediate gradients and avoids GPU memory limitations during contrastive training.

Experiments were conducted on NVIDIA H100 GPUs (80GB HBM3) using MIG partitions. Peak memory usage during training ranged from approximately 29GB to 40GB depending on the model architecture.

Training runtime varied across architectures due to differences in model size and interaction complexity. BioLORD fine-tuning on the Tier-1 completed in approximately 7.9 minutes (475 seconds), corresponding to roughly 460 optimization steps. Under unified ontology supervision, BioLORD training required approximately 1.53 hours (5,507 seconds) and approximately 3.7k optimization steps.

Qwen3-Embedding-0.6B required approximately 12.5 minutes (754 seconds) and 100 optimization steps for Tier-1 training, while unified supervision required approximately 1.50 hours (5,425 seconds) and 1,772 optimization steps.

Late-interaction reranking models based on GTE-ModernColBERT required longer training due to token-level interaction computation. Fine-tuning ModernColBERT on Tier-1A required approximately 55 minutes (3,304 seconds), corresponding to roughly 460 optimization steps, while unified ontology supervision required approximately 2.17 hours (7,818 seconds) and approximately 3.7k optimization steps.

Training throughput ranged from approximately

0.13 to 0.98 optimization steps per second across architectures. Learning rates were tuned per architecture:

- BioLORD-0.1B: 2×10^{-5}
- Qwen3-Embed-0.6B: 5×10^{-5}
- ModernColBERT: 3×10^{-6}

All other training parameters were held constant across supervision regimes to ensure that observed performance differences arise from supervision structure rather than optimization configuration.

E Evaluation Protocol

For each tier, queries rank all documents within the corresponding corpus without heuristic filtering or ontology pruning.

Metrics:

- Mean Reciprocal Rank (MRR)
- Recall@k (k = 1,5,10)
- nDCG@k (k = 1,5,10)

Metrics are computed independently per tier and averaged across queries.

F Prompt Used for Generative Candidate Scoring

We use the following prompt template for the Qwen3-4B-Think and Instruct models to evaluate the compatibility between a candidate disorder and a set of patient phenotypes.

System Instruction
You are a clinical expert evaluating whether a candidate disorder explains a patient's phenotypes.
Use ONLY the information provided in the phenotype list and the disorder description. Do NOT use external medical knowledge.

Patient phenotypes

- Increased lower facial height
- Laryngeal stenosis
- Long philtrum

Candidate disorder description
Autosomal dominant prognathism. Malocclusion in which the mandible is anterior to the maxilla as reflected by the first relationship of the first permanent molar.

Task
1. Identify phenotypes that **support** the disorder 2. Identify phenotypes that **contradict** the disorder 3. Assign a diagnostic compatibility score

Scoring guideline
0–20 incompatible with the phenotype set 20–40 weak overlap 40–60 partial match 60–80 strong match 80–100 near diagnostic match

Output format

```
{
  "score": integer (0-100),
  "supporting_features": [],
  "contradicting_features": []
}
```

G Additional Evaluation Metrics

Model	nDCG@k								
	Gene→Def			Phen→Def			Dis→Def		
	@1	@5	@10	@1	@5	@10	@1	@5	@10
BM25	0.079	0.083	0.087	0.265	0.361	0.381	0.200	0.262	0.275
MiniLM-L6-v2	0.158	0.244	0.263	0.500	0.609	0.629	0.224	0.296	0.314
BioLORD-0.1B‡	0.311	0.428	0.455	0.452	0.639	0.668	0.389	0.505	0.532
BioLORD-0.1B†	0.364	0.489	0.513	**0.767**	**0.862**	**0.869**	**0.425**	**0.540**	**0.562**
BioLORD-0.1B	0.167	0.268	0.290	0.693	0.795	0.807	0.309	0.402	0.426
MedEmbed-0.1B	0.320	0.424	0.440	0.600	0.722	0.738	0.350	0.433	0.454
SapBERT-0.1B	0.510	0.654	0.670	0.699	0.806	0.818	0.353	0.452	0.476
Qwen3-Embed-0.6B‡	0.183	0.298	0.331	0.318	0.464	0.498	0.279	0.401	0.434
Qwen3-Embed-0.6B†	0.284	0.431	0.459	0.648	0.782	0.796	0.336	0.444	0.470
Qwen3-Embed-0.6B	0.125	0.178	0.194	0.317	0.425	0.449	0.118	0.162	0.174
Qwen3-Embed-4B	0.228	0.329	0.354	0.319	0.429	0.457	0.118	0.183	0.203
OpenAI (text-embedding-3-large)	**0.619**	**0.730**	**0.747**	0.751	0.851	0.860	0.391	0.493	0.516

Table 6: Tier-1 ontology grounding tasks evaluated using nDCG@k.

Model	MRR@k								
	Gene→Def			Phen→Def			Dis→Def		
	@1	@5	@10	@1	@5	@10	@1	@5	@10
BM25	0.079	0.082	0.083	0.265	0.333	0.341	0.200	0.244	0.250
MiniLM-L6-v2	0.158	0.219	0.227	0.500	0.579	0.587	0.224	0.275	0.282
BioLORD-0.1B‡	0.311	0.392	0.404	0.452	0.586	0.598	0.389	0.472	0.483
BioLORD-0.1B†	0.364	0.454	0.464	**0.767**	**0.836**	**0.839**	**0.425**	**0.507**	**0.516**
BioLORD-0.1B	0.167	0.237	0.246	0.693	0.766	0.771	0.309	0.375	0.385
MedEmbed-0.1B	0.320	0.395	0.402	0.600	0.688	0.695	0.350	0.409	0.418
SapBERT-0.1B	0.510	0.614	0.620	0.699	0.777	0.782	0.353	0.423	0.433
Qwen3-Embed-0.6B‡	0.183	0.263	0.276	0.318	0.421	0.435	0.279	0.364	0.378
Qwen3-Embed-0.6B†	0.284	0.387	0.398	0.648	0.745	0.751	0.336	0.412	0.423
Qwen3-Embed-0.6B	0.125	0.162	0.168	0.317	0.394	0.403	0.118	0.149	0.154
Qwen3-Embed-4B	0.228	0.299	0.309	0.319	0.397	0.408	0.118	0.163	0.172
OpenAI (text-embedding-3-large)	**0.619**	**0.699**	**0.706**	0.751	0.825	0.829	0.391	0.464	0.473

Table 7: Tier-1 ontology grounding tasks. Metrics reported are MRR@k.

Model	Recall@k								
	Gene→Def			Phen→Def			Dis→Def		
	@1	@5	@10	@1	@5	@10	@1	@5	@10
BM25	0.079	0.088	0.100	0.265	0.448	0.508	0.200	0.315	0.357
MiniLM-L6-v2	0.158	0.318	0.378	0.500	0.700	0.761	0.224	0.360	0.417
BioLORD-0.1B‡	0.311	0.534	0.619	0.452	0.797	0.886	0.389	0.605	0.686
BioLORD-0.1B†	0.364	0.592	0.666	**0.767**	**0.937**	**0.958**	**0.425**	**0.639**	**0.706**
BioLORD-0.1B	0.167	0.362	0.429	0.693	0.879	0.916	0.309	0.485	0.559
MedEmbed-0.1B	0.320	0.508	0.559	0.600	0.821	0.873	0.350	0.505	0.569
SapBERT-0.1B	0.510	0.775	0.821	0.699	0.892	0.929	0.353	0.538	0.612
Qwen3-Embed-0.6B‡	0.183	0.405	0.507	0.318	0.594	0.699	0.279	0.511	0.612
Qwen3-Embed-0.6B†	0.284	0.565	0.651	0.648	0.890	0.933	0.336	0.539	0.621
Qwen3-Embed-0.6B	0.125	0.228	0.275	0.317	0.519	0.592	0.118	0.201	0.241
Qwen3-Embed-4B	0.228	0.421	0.497	0.319	0.526	0.611	0.118	0.242	0.306
OpenAI (text-embedding-3-large)	**0.619**	**0.824**	**0.875**	0.751	0.928	0.956	0.391	0.582	0.652

Table 8: Tier-1 ontology grounding tasks evaluated using Recall@k.

Model	nDCG@k											
	Dis→Gene			Dis→Phen			Phen→Dis			Phen→Gene		
	@1	@5	@10	@1	@5	@10	@1	@5	@10	@1	@5	@10
BM25	0.076	0.118	0.140	0.084	0.136	0.155	0.099	0.161	0.178	0.048	0.065	0.071
MiniLM-L6-v2	0.079	0.126	0.146	0.120	0.198	0.227	0.150	0.224	0.249	0.058	0.095	0.109
BioLORD-0.1B‡	0.112	0.186	0.213	0.158	0.289	0.333	0.034	0.182	0.250	**0.104**	**0.165**	**0.191**
BioLORD-0.1B†	0.061	0.103	0.123	0.130	0.215	0.245	0.151	0.248	0.276	0.054	0.078	0.091
BioLORD-0.1B	0.066	0.120	0.143	0.099	0.175	0.201	0.158	0.258	0.285	0.060	0.094	0.105
MedEmbed-0.1B	0.148	0.238	0.269	0.155	0.243	0.268	0.079	0.118	0.134	0.095	0.150	0.171
SapBERT-0.1B	0.104	0.168	0.194	0.142	0.232	0.257	0.068	0.101	0.113	0.096	0.146	0.168
Qwen3-Embed-0.6B‡	**0.332**	**0.515**	**0.546**	**0.371**	**0.561**	**0.586**	**0.244**	**0.383**	**0.414**	0.092	0.156	0.183
Qwen3-Embed-0.6B†	0.099	0.160	0.180	0.216	0.358	0.392	0.144	0.227	0.250	0.029	0.041	0.046
Qwen3-Embed-0.6B	0.169	0.241	0.264	0.209	0.316	0.343	0.155	0.236	0.257	0.034	0.051	0.058
Qwen3-Embed-4B	0.182	0.269	0.291	0.264	0.405	0.433	0.188	0.281	0.305	0.041	0.062	0.071
OpenAI (text-embedding-3-large)	0.092	0.149	0.170	0.140	0.223	0.248	0.176	0.276	0.303	0.084	0.121	0.136

Table 9: Tier-2 ontology relation tasks. Metrics reported are nDCG@k.

Model	MRR@k											
	Dis→Gene			Dis→Phen			Phen→Dis			Phen→Gene		
	@1	@5	@10	@1	@5	@10	@1	@5	@10	@1	@5	@10
BM25	0.076	0.105	0.114	0.084	0.120	0.128	0.099	0.143	0.150	0.048	0.060	0.062
MiniLM-L6-v2	0.079	0.112	0.120	0.120	0.175	0.187	0.150	0.202	0.214	0.058	0.084	0.090
BioLORD-0.1B‡	0.112	0.164	0.177	0.158	0.249	0.269	0.034	0.125	0.150	**0.104**	**0.146**	**0.156**
BioLORD-0.1B†	0.061	0.090	0.099	0.130	0.189	0.202	0.151	0.219	0.232	0.054	0.070	0.075
BioLORD-0.1B	0.066	0.103	0.113	0.099	0.152	0.163	0.158	0.228	0.241	0.060	0.084	0.088
MedEmbed-0.1B	0.148	0.211	0.225	0.155	0.217	0.229	0.079	0.106	0.113	0.095	0.134	0.143
SapBERT-0.1B	0.104	0.149	0.160	0.142	0.206	0.217	0.068	0.091	0.096	0.096	0.131	0.141
Qwen3-Embed-0.6B‡	**0.332**	**0.466**	**0.484**	**0.371**	**0.511**	**0.527**	**0.244**	**0.344**	**0.358**	0.092	0.136	0.147
Qwen3-Embed-0.6B†	0.099	0.142	0.151	0.216	0.317	0.334	0.144	0.202	0.213	0.029	0.038	0.040
Qwen3-Embed-0.6B	0.169	0.221	0.232	0.209	0.285	0.299	0.155	0.212	0.222	0.034	0.045	0.048
Qwen3-Embed-4B	0.182	0.245	0.256	0.264	0.366	0.380	0.188	0.255	0.266	0.041	0.056	0.059
OpenAI (text-embedding-3-large)	0.092	0.132	0.141	0.140	0.198	0.209	0.176	0.247	0.259	0.084	0.110	0.116

Table 10: Tier-2 ontology relation tasks. Metrics reported are MRR@k.

Model	Recall@k											
	Dis→Gene			Dis→Phen			Phen→Dis			Phen→Gene		
	@1	@5	@10	@1	@5	@10	@1	@5	@10	@1	@5	@10
BM25	0.076	0.160	0.232	0.084	0.184	0.245	0.099	0.218	0.277	0.048	0.081	0.100
MiniLM-L6-v2	0.079	0.172	0.236	0.120	0.273	0.369	0.150	0.295	0.379	0.058	0.128	0.173
BioLORD-0.1B‡	0.112	0.258	0.349	0.158	0.419	0.570	0.034	0.342	0.524	**0.104**	**0.225**	**0.306**
BioLORD-0.1B†	0.061	0.145	0.213	0.130	0.298	0.397	0.151	0.341	0.435	0.054	0.104	0.143
BioLORD-0.1B	0.066	0.171	0.248	0.099	0.248	0.334	0.158	0.354	0.449	0.060	0.124	0.159
MedEmbed-0.1B	0.148	0.326	0.429	0.155	0.328	0.415	0.079	0.155	0.204	0.095	0.202	0.274
SapBERT-0.1B	0.104	0.232	0.314	0.142	0.319	0.401	0.068	0.131	0.169	0.096	0.197	0.272
Qwen3-Embed-0.6B‡	**0.332**	**0.701**	**0.834**	**0.371**	**0.745**	**0.864**	**0.244**	**0.512**	**0.620**	0.092	0.217	0.300
Qwen3-Embed-0.6B†	0.099	0.217	0.287	0.216	0.492	0.615	0.144	0.305	0.388	0.029	0.053	0.067
Qwen3-Embed-0.6B	0.169	0.312	0.396	0.209	0.421	0.522	0.155	0.313	0.390	0.034	0.068	0.089
Qwen3-Embed-4B	0.182	0.353	0.436	0.264	0.538	0.649	0.188	0.370	0.455	0.041	0.082	0.109
OpenAI (text-embedding-3-large)	0.092	0.204	0.277	0.140	0.302	0.387	0.176	0.372	0.466	0.084	0.157	0.202

Table 11: Tier-2 ontology relation tasks. Metrics reported are Recall@k

Model	nDCG@k		
	PhenTriplet→Dis		
	@1	@5	@10
BM25	0.066	0.120	0.138
MiniLM-L6-v2	0.072	0.127	0.148
BioLORD-0.1B‡	0.085	0.231	0.283
BioLORD-0.1B†	0.096	0.174	0.204
BioLORD-0.1B	0.082	0.154	0.182
MedEmbed-0.1B	0.083	0.123	0.134
SapBERT-0.1B	0.076	0.116	0.127
Qwen3-Embed-0.6B‡	**0.182**	**0.307**	**0.344**
Qwen3-Embed-0.6B†	0.064	0.115	0.138
Qwen3-Embed-0.6B	0.044	0.082	0.098
Qwen3-Embed-4B	0.020	0.049	0.064
OpenAI (text-embedding-3-large)	0.097	0.172	0.200

Table 12: Tier-3 compositional retrieval task. Metrics reported are nDCG@k.

Model	MRR@k		
	PhenTriplet→Dis		
	@1	@5	@10
BM25	0.066	0.103	0.110
MiniLM-L6-v2	0.072	0.110	0.119
BioLORD-0.1B‡	0.085	0.185	0.207
BioLORD-0.1B†	0.096	0.151	0.163
BioLORD-0.1B	0.082	0.131	0.143
MedEmbed-0.1B	0.083	0.123	0.134
SapBERT-0.1B	0.076	0.116	0.127
Qwen3-Embed-0.6B‡	**0.182**	**0.270**	**0.286**
Qwen3-Embed-0.6B†	0.064	0.099	0.109
Qwen3-Embed-0.6B	0.044	0.070	0.077
Qwen3-Embed-4B	0.020	0.040	0.046
OpenAI (text-embedding-3-large)	0.097	0.149	0.160

Table 13: Tier-3 compositional retrieval task. Metrics reported are MRR@k.

Model	Recall@k		
	PhenTriplet→Dis		
	@1	@5	@10
BM25	0.066	0.172	0.230
MiniLM-L6-v2	0.072	0.180	0.246
BioLORD-0.1B‡	0.085	0.368	0.527
BioLORD-0.1B†	0.096	0.247	0.340
BioLORD-0.1B	0.082	0.224	0.312
MedEmbed-0.1B	0.083	0.123	0.281
SapBERT-0.1B	0.076	0.116	0.272
Qwen3-Embed-0.6B‡	**0.182**	**0.425**	**0.547**
Qwen3-Embed-0.6B†	0.064	0.165	0.236
Qwen3-Embed-0.6B	0.044	0.117	0.170
Qwen3-Embed-4B	0.020	0.077	0.126
OpenAI (text-embedding-3-large)	0.097	0.244	0.331

Table 14: Tier-3 compositional retrieval task. Metrics reported are Recall@k.

Task	Method	MRR@1	Recall@1	nDCG@1	MRR@5	Recall@5	nDCG@5	MRR@10	Recall@10	nDCG@10
1B-R (Dis→Phen)	Phenomizer-style	0.622	0.622	0.622	0.704	0.834	0.721	0.711	0.885	0.716
1C-R (Phen→Dis)	Phenomizer-style	0.713	0.713	0.713	0.795	0.916	0.823	0.800	0.949	0.831
3A (PhenTriplet→Dis)	Phenomizer-style	0.635	0.635	0.635	0.729	0.880	0.764	0.737	0.937	0.778
2A-R (Phen→Gene)	Graph Traversal	0.799	0.799	0.799	0.830	0.867	0.840	0.830	0.867	0.840
2B-R (Dis→Gene)	Graph Traversal	0.933	0.933	0.933	0.933	0.933	0.933	0.933	0.933	0.933

Table 15: Performance of ontology-based baselines across tasks. Phenotype–disease and compositional tasks use a Phenomizer-style ontology-based similarity, while gene-related tasks use graph traversal based on disease–gene connectivity.

H Error Analysis (Tier3)

Failure Mode Definition and Assignment We perform a systematic, rule-based error analysis on Tier-3 cases where the embedding retriever fails but ontology-aware baselines succeed, isolating residual errors under controlled conditions.

Signals. For each query, we compute: (i) phenotype overlap between the query and predicted disease (Ov), (ii) ontology-based similarity between predicted and true disease, and (iii) average information content (IC) of phenotypes associated with the predicted disease.

Assignment rules. Failure modes are assigned deterministically in the following order:

- **Related-disease confusion:** The predicted disease is highly similar to the true disease (Phenomizer-style baseline $\geq \tau_{\text{high}}$), but incorrect.
- **Generic bias:** The predicted disease is characterized by low-specificity phenotypes (average IC $\leq \tau_{\text{IC}}$).
- **Aggregation failure:** The predicted disease matches at most one query phenotype (Ov ≤ 1), indicating failure to combine phenotype evidence.
- **Semantic drift:** Remaining cases with low phenotype overlap and low ontology similarity.

Notes. Given that Tier-3 queries consist primarily of phenotype triplets, we distinguish strict aggregation failures (Ov ≤ 1) from near-miss cases (Ov $= 2$), which are resolved by ontology similarity. All categories are mutually exclusive by construction.

Mode	QID	Phenotypes	Predicted disease (embedding model)	True disease	$r_{\text{Phenomizer}}$	Ov_{pred}	Ov_{true}
Aggregation	Q2599	Sloping forehead; 2–4 toe cutaneous syndactyly; Chiari type II malformation	Crossed polysyndactyly	Lathosterolosis	1	0	3
Aggregation	Q4636	Finger syndactyly; Ungual fibroma; Glue ear	Crossed polysyndactyly	Choroidal atrophy-alopecia syndrome	1	1	3
Aggregation	Q5613	Long philtrum; Conical incisor; Hypophosphaturia	Hypophosphatemic rickets, autosomal recessive, 2	Global developmental delay-osteopenia-ectodermal defect syndrome	1	0	3
Related confusion	Q4072	Flat face; Sacral dimple; Prominent fingertip pads	PDE4D haploinsufficiency syndrome	Vulto-van Silfout-de Vries syndrome	1	2	3
Related confusion	Q1943	Short philtrum; Prominent fingertip pads; Scoliosis	1q21.1 microdeletion syndrome	Global developmental delay-visual anomalies-progressive cerebellar atrophy-truncal hypotonia syndrome	1	2	3
Related confusion	Q2002	Midface retrusion; Scoliosis; Radial deviation of finger	Autosomal dominant Robinow syndrome	Acrofacial dysostosis 1, Nager type	1	2	3
Generic	Q18	Intellectual disability; Tremor; Tapered finger	Developmental delay and seizures with or without movement abnormalities	Agenesis of the corpus callosum with peripheral neuropathy	1	2	3
Generic	Q976	Abnormal cardiovascular system morphology; Long philtrum; Frontal bossing	1q21.1 microdeletion syndrome	Doors syndrome	1	2	3
Generic	Q730	Seizure; Splenomegaly; Ptosis	Sialidosis type 2	Proteus syndrome, somatic	1	2	3
Semantic drift	Q6075	Ptosis; Paraplegia; Brain atrophy	Primary lateral sclerosis	Oculopharyngodistal myopathy 1	12	0	3
Semantic drift	Q651	Brachydactyly; Limitation of joint mobility; Abnormal metaphysis morphology	Metaphyseal dysostosis-intellectual disability-conductive deafness syndrome	Thanatophoric dysplasia type 2	12	0	3
Semantic drift	Q2375	Generalized-onset seizure; Lower limb spasticity; Status epilepticus	Epileptic encephalopathy, early infantile, 6 (Dravet syndrome)	Multiple congenital anomalies-hypotonia-seizures syndrome 2	13	1	3

Table 16: Representative Tier-3 embedding failures grouped by rule-based error mode. The embedding model frequently retrieves partially compatible disorders despite full phenotype coverage by the ground-truth disease, while the Phenomizer baseline often resolves the same queries through ontology-based phenotype aggregation. QID denotes the query identifier. $r_{\text{Phenomizer}}$ denotes the ranking position assigned to the ground-truth disease by the ontology-aware Phenomizer baseline. Ov_{pred} and Ov_{true} denote phenotype overlap counts between the query phenotype set and the predicted or ground-truth disease, respectively.

H.1 Tier-3 Failure Summary: Base vs. OntologyBench Qwen3-0.6B

Metric	Base Qwen	OntologyBench FT
Ground-truth disorder not retrieved (%)	70.6%	29.4%
Mean rank of retrieved ground-truth disorders	26.47	23.33
Top-1 disagreement rate vs ontology baseline	94.4%	75.8%

Table 17: Tier-3 failure summary comparing the base and OntologyBench fine-tuned Qwen3-0.6B models. Fine-tuning substantially reduces catastrophic retrieval failures where the ground-truth disorder is absent from retrieval and improves agreement with ontology-aware retrieval behavior.