Bridging Lexical Divergence: LLM-Assisted, Cost-Efficient, Zero-shot Scientific Entity Linking

Md Rasel Khondokar*, **Qiao Qiao***, **Farjana Sultana Samia, Nhat Le, Yuepei Li, Qi Li**
Department of Computer Science, Iowa State University, Ames, Iowa, USA
{rasel, qqiao1, fssamia, lan0908, liyp0095, qli}@iastate.edu

Abstract

Scientific domain entity linking (EL) differs from general domain EL because mentions and entity names often lack lexical overlap. Another challenge is that specialized terminology is used in the scientific domain, which is rarely encountered in models pretrained on general domains. Therefore, models trained on general domains transfer poorly to scientific domains. To address this, in-domain fine-tuning is the natural remedy. However, many scientific domains lack expert-annotated data, motivating the need for a zero-human-annotation approach. Existing zero-shot methods heavily rely on LLMs to generate aliases across entire mention corpora, which incurs substantial computational cost, and those methods provide no mechanism to filter out noise from LLMs. To address these challenges, we propose Sci-ZSEL[1], a framework that selectively generates entity aliases with an LLM to control computational cost, and applies an ontology-aware filter to remove aliases that semantically drift toward ontology neighbors. Then, filtered aliases are used to construct pseudo-labeled mention-entity pairs for fine-tuning. To enable evaluation of EL under low lexical overlap, we also release a new animal science EL benchmark linked to three livestock trait ontologies, where mentions and entities exhibit substantially lower lexical overlap than in existing benchmarks. Across five benchmarks, Sci-ZSEL outperforms the non-fine-tuned baseline, is most useful on non-overlapping mentions, and combining it with curated synonyms gives the best performance in most settings.

1 Introduction

Entity Linking (EL) identifies mentions in text and links them to entities in a **Knowledge Graph (KG)**. Most EL research targets **general domain EL**, where the central challenge is **disambiguation** among candidates with similar surface forms (e.g., "Apple" the company vs. the fruit). **Zero-Shot Entity Linking (ZSEL)** aims to perform linking without requiring human-annotated training data. Recent ZSEL approaches (Wu et al., 2020; Xu et al., 2023b; Zhou et al., 2024) have made significant progress by leveraging entity descriptions and contextual embeddings to resolve such ambiguities. However, these methods are typically trained on open-domain corpora such as Wikipedia, where mentions and entities often exhibit high lexical overlap.

Scientific domain EL faces a different challenge from general domain EL, where mentions and entity names often **lack lexical overlap**. Rather than resolving ambiguity among lexically similar candidates, scientific EL must handle highly variable terminology, where formal names may differ significantly from their common or colloquial names. For example, a paper may use “lambing potential” instead of “goat fertility”, and “hypertension” instead of “high blood pressure”. This **lexical divergence** poses a major obstacle for standard approaches, which rely on token-level similarity. Another unique characteristic of scientific EL is that the KG is often represented as an ontology, with terms organized in hierarchical structures that show relationships between concepts. Such information is not sufficiently leveraged in existing ZSEL models trained on general domain corpora.

Consequently, pretrained ZSEL models struggle with scientific EL tasks. First, due to differences in lexical challenges between general and specialized scientific domains, they struggle to transfer effectively. Second, they cannot handle specialized terminology that the general domain rarely encounters. Third, pretrained ZSEL models ignore the ontology structure, treating each entity as an isolated string and missing the hierarchical signal that distinguishes closely related concepts. All limita-

*Equal contribution.
[1]Code and data: https://github.com/rasel-isu/Sci-ZSEL

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 27231–27249
October 24-29, 2026 ©2026 Association for Computational Linguistics

tions point to fine-tuning on in-domain data as the natural remedy, and prior work has taken this route (Yuan et al., 2022; Xu et al., 2023a). However, fine-tuning requires labeled mention-entity pairs, and expert annotation is expensive and unavailable in many specialized domains. Recent work uses LLMs to synthesize training data (Xin et al., 2025; Ye and Mitchell, 2025). Still, these methods use an LLM for each mention: the token cost scales with the size of the corpus, and the generated pairs are tied to that corpus rather than being reusable across new corpora drawn from the same ontology.

To address these challenges, we propose Sci-ZSEL, a cost-aware framework that does not require human-labeled mention-entity pairs from the target domain. The framework generates pseudo-labeled training pairs by prompting an LLM for alias names that serve as lexical bridges between corpus mentions and ontology entities. To bound LLM cost, we apply this augmentation on the entity side rather than the mention side: instead of issuing one LLM call per corpus mention, we identify a small set of entities and generate aliases for each, then use those aliases to find mentions to form mention-entity pairs for fine-tuning. To ensure that LLM-generated aliases properly align with the entities in the ontology, we further apply an ontology-aware filter that uses entity hierarchy and semantic similarity to discard drifted aliases before training.

We further release a new EL benchmark drawn from animal science literature. This dataset is curated by domain experts and highlights the lexical divergence challenge in scientific EL tasks.

Our main contributions are as follows:

- A new animal science EL benchmark dataset formed based on PubMed articles linked to three livestock trait ontologies.
- **Sci-ZSEL**, a zero-shot EL framework that uses an LLM to generate alias names for selected entities and builds pseudo pairs from real mentions and original entity names.
- Sci-ZSEL improves over the non-fine-tuned baseline across all five benchmarks; moreover, combining Sci-ZSEL generated aliases with curated synonyms achieves the best performance in most settings.

2 Related Work

General domain EL primarily targets *disambiguation* among lexically similar candidates. Early systems exploited surface overlap and context heuristics (Cucerzan, 2007; Ratinov et al., 2011), and later methods used pretrained language models and introduced dense semantic representations that improved robustness to surface variation (Yamada et al., 2016; Gillick et al., 2019) where they assume substantial lexical overlap between mention and entity, which is an assumption that fails in scientific text.

2.1 Zero-Shot Entity Linking

ZSEL removes dependence on in-domain labelled mentions by linking through entity descriptions that enable generalization to unseen entities (Logeswaran et al., 2019; Wu et al., 2020; Xu et al., 2023b; Zhou et al., 2024). Existing ZSEL methods show strong performance on Wikipedia-style benchmarks where mentions and entity names have lexical overlap, but they implicitly rely on surface-form similarity and transfer poorly to specialized domains where mentions and entities may share no lexical overlap at all. Generative EL systems, such as GENRE (De Cao et al., 2021), struggle with unseen entities in the zero-shot setting.

2.2 Biomedical Entity Linking

Biomedical EL faces this lexical divergence most acutely: abbreviations, acronyms, and synonyms produce mention-entity pairs with minimal surface overlap (e.g., "EGFR" vs. "Epidermal Growth Factor Receptor"). Early dictionary and rule-based methods offer high precision but limited scalability (Aronson, 2001; Leaman et al., 2013). Neural approaches with domain-specific encoders (Sung et al., 2020; Liu et al., 2021; Chen et al., 2021), generative formulations (Yuan et al., 2022), and cross-entity interaction (Xu et al., 2023a; Kim et al., 2025) can improve performance but still depend on large expert-annotated corpora for fine-tuning, which are unavailable in emerging scientific domains where ontology synonyms are also sparse.

2.3 LLM-Based Data Augmentation

Recent work uses LLMs to synthesize EL supervision, either by augmenting mention contexts or by acting as end-to-end disambiguators (Xin et al., 2025; Sanz-Cruzado and Lever, 2025; Ye and Mitchell, 2025). These approaches generate per-mention, so token cost scales with corpus size, and there is no guard against drift. Sci-ZSEL inverts the direction: it generates aliases for a subset of entities, pairs them with observed corpus mentions,

and applies a filter to discard drifted aliases, which combines the semantic reach of LLMs with the efficiency of a compact retriever-reranker pipeline.

3 Animal Science Benchmark

Animal science research on quantitative trait loci (QTL) requires normalizing free-text trait mentions to enable combining findings across studies. Trait names, however, differ across research aspects, species, and product lines, and are therefore curated in separate ontologies.

In the proposed animal science benchmark, the EL task utilizes three livestock trait ontologies, Clinical Measurement Ontology (CMO), Vertebrate Trait Ontology (VT), and Livestock Product Trait Ontology (LPT), because they are used widely for animal QTL studies. The EL task expects recognised entities to perform linking, and in order to follow the zero-shot setting, we do not annotate a training set. Instead, we use the AnimalQTLdb Benchmark released in Hu et al. (2007) as the training corpus and use the CuPUL method (Li et al., 2025) to recognize trait mentions.

The test dataset is annotated manually using two strategies. First, a domain expert randomly selected 150 records from AnimalQTLdb, an animal science database curated for QTL research. Each record contains a normalized trait name, the ontology IDs from the aforementioned 3 ontologies, and the corresponding publication (PubMed ID) from which the record is curated. The expert then manually identifies corresponding mentions from the full paper. Second, we obtain a comprehensive list of trait names and their commonly observed mentions from AnimalQTLdb curators. Then, we collect all abstracts of papers curated in AnimalQTLdb and apply string matching to identify the trait mentions. Four student annotators manually identified the entities from the three ontologies and corrected mention boundaries when needed. Finally, the domain expert validated the annotations.

The statistics of two existing biomedical benchmarks NCBI Disease (Dogan et al., 2014), BC5CDR (Li et al., 2016), and the new animal science benchmarks (QTL_{CMO}, QTL_{VT}, and QTL_{LPT}) are summarized in Table 1.

To characterize different levels of lexical divergence, we partition mention–entity pairs into three categories based on their degree of lexical overlap. **HO** (high overlap) refers to cases where the mention is a substring of the entity name. **LO** (low overlap) refers to cases where the mention and entity name exhibit partial lexical overlap, but the mention is not a substring of the entity name. **NO** (no overlap) refers to cases where the mention and entity name share no lexical overlap.

The resulting test sets differ from existing biomedical benchmarks in two ways. First, the NO rate is substantially higher than existing benchmarks. Moreover, the high ratio of unique mentions to unique entities indicates that the mentions are highly diverse for the same entities. On the other hand, the three ontologies from the animal science domain are less developed than those in the biomedical domains, with a lower percentage of entities with synonyms and fewer synonyms per entity. Both observations illustrate that our new animal science benchmark may be significantly harder than the existing biomedical benchmarks.

Annotation Limitations: Due to the high cost of domain-expert review, the test sets are small, which limits statistical power for fine-grained comparisons. The pipeline may also carry a selection bias: we select mentions using dictionary string matching against a list of commonly observed mentions provided by domain experts, so mentions absent from those lists are dropped before annotators see them. However, the list is significantly larger than synonyms provided in ontologies and contains more diverse expressions. For example, “IMF” in this list is widely used to refer to "intramuscular fat" in QTL articles but does not appear in any ontologies. Further, annotators partially offset this by manually collecting no-overlap cases. However, the test mention pool remains biased toward surface forms that the ontologies anticipate.

4 Preliminaries

Let $\mathcal{E}$ be the set of entities in a target ontology and $\mathcal{M}$ the mentions extracted from an unlabeled training corpus. Each entity $e \in \mathcal{E}$ has a name $n(e)$, a definition $d(e)$, and may have a list of synonyms $\text{Syn}(e)$. Given a mention $m \in \mathcal{M}$ with context $c(m)$, the task is to predict its referent entity $e \in \mathcal{E}$. For the zero-shot setting, no human-labeled $\langle m, e \rangle$ pairs are available for training. In this paper, we use *zero-shot* to denote the zero-human-label setting.

A bi-encoder retriever maps the mention and the entity to dense vectors $\tau_m, \tau_e \in \mathbb{R}^d$ and scores them by their dot product $\tau_m \cdot \tau_e$. It returns the top-K entities under this score. We use the pretrained BLINK (Wu et al., 2020) as the bi-encoder

Dataset name	Ontology Name	Test set statistics						Ontology statistics			
		Samples	HO%	LO%	NO%	Ment	Ent	Entities	Synonyms	Syn/Ent	Cov%
NCBI Disease	MEDIC	960	32.08	45.00	22.92	287	190	13,316	127,370	9.57	85.0
BC5CDR	MeSH	9465	62.07	13.77	24.16	2084	1267	355,213	757,086	2.13	94.5
QTL_{CMO}	CMO	2032	20.13	36.42	43.45	458	107	4,133	4,413	1.07	57.2
QTL_{VT}	VT	1688	8.83	34.06	57.11	500	111	4,044	3,670	0.91	44.0
QTL_{LPT}	LPT	722	30.75	24.24	45.01	233	103	520	462	0.89	41.7

Table 1: Combined test-set and ontology statistics. Test set block: Samples refers to the number of mention-entity pairs in the test set, HO, LO, and NO show the percentages of samples by overlap categories; and Ment and Ent refer to the number of unique mentions and unique entities in the test set, respectively. Ontology block: Entities and Synonyms refers to the total number in the ontology, Syn/Ent refers to averaged synonyms per entity, and Cov% is the percentage of entities with at least one synonym provided. Note that QTL_{CMO}, QTL_{VT}, and QTL_{LPT} are test-only datasets.

backbone; details in Appendix A.3.

A cross-encoder reranker encodes mention m, its context $c(m)$, and all candidates e jointly to produce a score $s(m, e)$. We consider two pretrained cross-encoder backbones, BLINK (Wu et al., 2020) and ReS (Xu et al., 2023b); details in Appendix A.4.

5 Methodology

5.1 Entity Selection from the Ontology

An ontology defines the vocabulary of a domain and often contains a large collection of entities. However, following the assumption of Zipf's law (Zipf, 1949), many of those entities are not mentioned in the target corpus, so it is unnecessary to consider every entity. Therefore, to reduce cost, Sci-ZSEL strategically selects a subset of entities and uses LLM to generate their aliases. This also makes the generated supervision reusable across corpora collected from similar domains. Figure 1 shows the entire flow of pseudo pair generation.

In scientific domains, mention ambiguity is relatively low. If a mention matches the surface name of an entity, the pairing is usually correct. Following this assumption, we include these entities whose name appears as mentions in the corpus:

$$\mathcal{E}_{\text{EM}} = \left\{ e \in \mathcal{E} : \exists\, m \in \mathcal{M},\ m \equiv n(e) \right\}. \quad (1)$$

A pretrained BLINK bi-encoder embeds mentions and entities into a semantic space, so top-ranked entities can provide a semantic bridge for lexically divergent mentions. To do so, we pass each mention $m \in \mathcal{M}$ to the bi-encoder and use its top-1 entity. Specifically, we define this set as $\mathcal{E}_{\text{BT}}$:

$$\mathcal{E}_{\text{BT}} = \left\{ \text{top1}_{\text{BLINK}}(m) : m \in \mathcal{M} \right\}. \quad (2)$$

5.2 Pseudo-Pair Construction

We construct pseudo-labeled pairs by using the unlabeled training corpus and the selected entities ($\mathcal{E}_{\text{EM}} \cup \mathcal{E}_{\text{BT}}$) from Section 5.1 via the three construction strategies below.

5.2.1 Construction from Surface Name Matching

Similar to exact match entity selection, but here, we return $\langle m, e \rangle$ pair for the pseudo training data instead of entity. Mathematically, it is defined as:

$$\mathcal{P}_1 = \left\{ \langle m, e \rangle : m \in \mathcal{M},\ e \in \mathcal{E},\ m \equiv n(e) \right\}. \quad (3)$$

5.2.2 Construction from LLM-generated Alias

LLMs possess semantically related lexical expression capabilities; we leverage them to generate representative aliases from entity definitions. Our goal is to synthesize valid surface forms that mentions in the corpus might use. To do so, we prompt the LLM with the definition $d(e)$ of an entity $e \in \mathcal{E}_{\text{EM}} \cup \mathcal{E}_{\text{BT}}$ to produce aliases $a(e)$ (the prompt template is provided in Appendix A.1). We apply alias generation to all entities in $\mathcal{E}_{\text{EM}} \cup \mathcal{E}_{\text{BT}}$, producing the alias set $\mathcal{E}_{\text{A}}$.

Ontology-Aware Filtering. A common failure mode of LLM generation is semantic drift, where an alias inadvertently describes a hierarchical neighbor of e rather than the entity itself. To mitigate this, we filter the generated aliases $a(e)$ by comparing their semantic similarity to the target entity name $n(e)$ against the entity's ontology neighbors.

We first define the ontology neighborhood of entity e using its parents, children, and siblings:

$$\mathcal{N}(e) = \text{parents}(e) \cup \text{children}(e)\,, \quad (4)$$

$$\mathcal{N}^{+}(e) = \mathcal{N}(e) \cup \text{siblings}(e)\,. \quad (5)$$

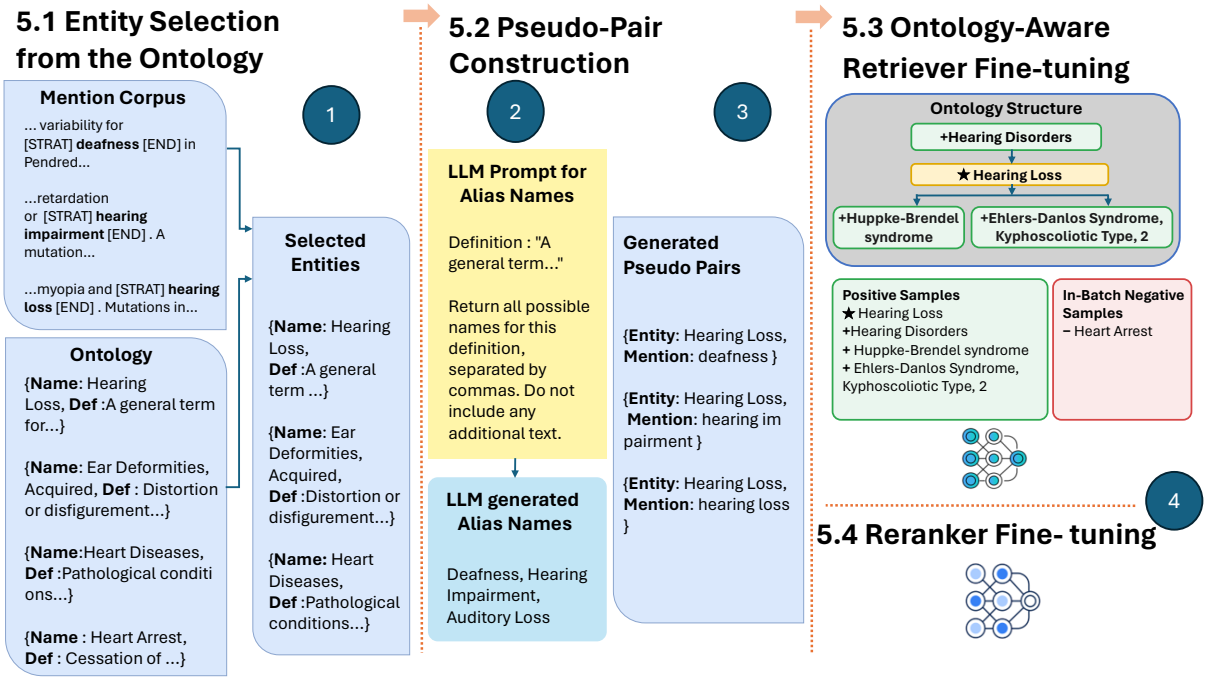


Figure 1: Overview of the Sci-ZSEL framework with a worked example. **(1)** Candidate entities are selected from the ontology using the unlabeled mention corpus. **(2)** Selected entity definitions are sent to the LLM to generate aliases. **(3)** The pseudo-pairs are used to fine-tune the retriever, where ontology neighbors of the gold entity (parents and children) are promoted as positives (★ gold entity, + positive, − in-batch negative) and other entities within the batch act as in-batch negatives. **(4)** The pseudo-pairs are used to fine-tune the reranker.

For each generated alias $a(e) \in \mathcal{E}_{\mathrm{A}}$, we compute an anchor similarity to the entity name $n(e)$ and a drift similarity to its most closely related neighbor in $\mathcal{N}^{+}(e)$:

$$s_{\text{anchor}} = \mathrm{sim}\big(n(e),\, a(e)\big), \tag{6}$$

$$s_{\text{drift}} = \max_{e_{\mathcal{N}} \in \mathcal{N}^{+}(e)} \mathrm{sim}\big(n(e_{\mathcal{N}}),\, a(e)\big). \tag{7}$$

where $\mathrm{sim}(\cdot,\cdot)$ denotes the cosine similarity between BioLORD (Remy et al., 2024) embeddings of two entity names.

We discard the alias $a(e)$ if $s_{\text{anchor}} < s_{\text{drift}}$ or $s_{\text{anchor}} < \tau$, where $\tau = 0.9$. The first condition removes aliases that are semantically closer to a neighboring entity than to the target entity e, while the second removes aliases that diverge excessively from the target entity name $n(e)$. The resulting filtered alias set is denoted as $\mathcal{E}_{\mathrm{AF}}$.

Pseudo-Label Construction. Finally, we search for the filtered aliases $a(e) \in \mathcal{E}_{\mathrm{AF}}$ within the unlabeled corpus $\mathcal{M}$. We form a pseudo-labeled pair whenever an observed mention m exactly matches a filtered alias $a(e)$ (denoted as $m \equiv a(e)$ after whitespace and case normalization). Specifically, we generate the following sets of pseudo-labeled pairs:

$$\mathcal{P}_2 = \{\langle m, e\rangle : e \in \mathcal{E}_{\mathrm{AF}}, m \in \mathcal{M}, m \equiv a(e)\}\,. \tag{8}$$

The combined set is:

$$\mathcal{P}_{\mathrm{Gen}} = \mathcal{P}_1 \cup \mathcal{P}_2\,. \tag{9}$$

5.2.3 Construction from Ontology Synonym

Many ontologies contain expert-curated synonym lists designed to capture alternative surface forms of an entity's name. Therefore, we can directly leverage these curated synonyms, denoted as $s \in \mathrm{Syn}(e)$, as highly reliable lexical bridges. We construct a pseudo-labeled pair for an entity e whenever one of its curated synonyms s exactly matches an observed mention m in the unlabeled corpus. Formally, it is defined as:

$$\mathcal{P}_{\mathrm{Syn}} = \big\{\, \langle m, e\rangle : e \in \mathcal{E},\;\; m \in \mathcal{M},\; m \equiv s \,\big\}\,. \tag{10}$$

5.3 Ontology-Aware Retriever Fine-tuning

The retriever determines which candidate entities are exposed to the reranker. We therefore fine-tune the retriever to better rank semantically related ontology entities. Hierarchical neighbors in an ontology often have related meanings; they provide informative training signals for improving retrieval

under the ontology structure. To do that, we incorporate ontology-aware sampling strategies to the retriever. The **BASE** strategy treats every non-gold entity as a negative. The **RM-PCS** removes neighbors $\mathcal{N}^+(e)$ from the negatives. The **PC-POS** promotes parents and children $\mathcal{N}(e)$ of e, as they are added as positives, while siblings remain as negatives.

Each strategy uses an in-batch cross-entropy objective with these introduced positive and negative sets; the goal is to score positive pairs higher than negative pairs within each batch. Full formulation is in Appendix A.3.2.

5.4 Reranker Fine-tuning

The primary responsibility of the reranker is to learn the accurate distinctions among relevant top-K candidates, then assign the highest score to the most accurate candidate. Therefore, we do not apply the ontology-aware sampling strategies here. Instead, we fine-tune the BLINK cross-encoder and ReS using their standard architecture (Section 4).

For each pseudo pair $\langle m, e \rangle$, we retrieve the top-K candidates from the bi-encoder, treat e as the positive and randomly selected candidates as negatives from the rest, then optimize a cross-entropy loss over the candidate set. The goal is to rank the pseudo-labeled entity at the top. Full formulation and the candidate-set augmentation rule (when e is missing from top-K) are in Appendix A.4.

6 Experimental Studies

6.1 Datasets

We evaluate Sci-ZSEL on five benchmarks. NCBI Disease (Dogan et al., 2014) (linking to MEDIC (Davis et al., 2012)) and BC5CDR (Li et al., 2016) (linking to MeSH (Lipscomb, 2000)) are existing synonym-rich biomedical benchmarks; QTL_{CMO}, QTL_{VT}, and QTL_{LPT} are synonym-sparse animal science benchmarks released with this paper (Section 3). Training pairs are constructed as described in Section 5.2. Statistics for test sets and training pairs are in Table 1 and Table 7, respectively.

6.2 Baselines

We select baseline backbones that satisfy two requirements. First, a baseline must be a zero-shot method or otherwise applicable to an out-of-domain setting. Second, its pretrained model must be publicly accessible so that it can be directly used as a backbone in our framework.

Therefore, we instantiate Sci-ZSEL on two pretrained general-domain ZSEL backbones: BLINK (Wu et al., 2020) for the retriever and reranker, and ReS (Xu et al., 2023b) for the reranker. For each backbone on each dataset, we run five fine-tuning settings to isolate the contributions. The **No fine-tune** setting uses the frozen, pretrained ZSEL backbone to exhibit the performance under scientific domain shift. **Sci-ZSEL w/o filter** fine-tunes on the unfiltered LLM-generated alias setting (omitting the step from Section 5.2.2) to see the raw impact. **Sci-ZSEL** fine-tunes after filtering, so it uses $\mathcal{P}_{\text{Gen}}$ (Eq. 9), which demonstrates the impact and necessity of the ontology-aware filter. **Synonym** uses $\mathcal{P}_{\text{Syn}}$ (Eq. 10) that exhibits whether curated synonyms alone suffice for EL, where they are abundant, and **Sci-ZSEL + Synonym** $\mathcal{P}_{\text{Gen}} \cup \mathcal{P}_{\text{Syn}}$ shows whether LLM aliases and curated synonyms are complementary.

6.3 Experimental Setup

We fine-tune two pretrained zero-shot EL backbones, BLINK and ReS, using their original architectures. In a zero-shot setting, we do not have a validation set, so best-epoch selection is not possible; we therefore fix the number of epochs and report the results from the last epoch. All reported results are means over three random seeds (0, 42, 52313). Details in Appendix A.5.

6.4 Evaluation Metrics

We evaluate the retriever using Recall@64 and the reranker using Recall@1. Recall@K measures the proportion of mentions for which the gold entity appears among the top-K candidates. We further report mean reciprocal rank (MRR), which measures the average reciprocal rank of the gold entity across test mentions, as well as results on the HO, LO, and NO subsets defined in Section 3 to assess performance under varying levels of lexical divergence.

6.5 Experimental Results

In this section, we begin by analyzing the quality of our pseudo training pairs in Section 6.5.1. Next, we evaluate the retriever's performance, presenting its overall Recall@64 scores in Table 4 and providing MRR and category-wise performance (HO, LO, and NO) in Tables 10, 11, and 12. Then, we evaluate the reranker's performance in

Ontology	Exact match ($\mathcal{E}_{\text{EM}}$)			Biencoder top-1 ($\mathcal{E}_{\text{BT}}$)		
	Ent	%Syn	%Name	Ent	%Syn	%Name
MEDIC	169	26.04	54.44	636	29.87	29.09
MeSH	721	10.54	53.68	1341	8.58	41.83
CMO	58	5.17	36.21	1260	3.33	10.00
VT	20	0.00	40.00	1396	2.22	7.88
LPT	30	3.33	46.67	361	2.22	14.68

Table 2: LLM-generated alias behavior across source ontologies. Ent is the number of entities used for alias generation; %Syn and %Name are the percentages of aliases matching a curated synonym and the original entity name, respectively (the remainder are new lexical bridges).

Table 5, highlighting its overall Recall@1, MRR, and category-wise results. To ensure our findings are stable and reliable, every number we report is an average calculated across three separate test runs using random seeds 0, 42, and 52313.

Dataset	Setting	Accuracy	pseudo pair count
NCBI	P_{Gen} W/O filter	72.91	1757
	$\mathcal{P}_{\text{Gen}}$	**90.32**	1301
BC5CDR	P_{Gen} W/O filter	72.55	7447
	$\mathcal{P}_{\text{Gen}}$	**91.4**	5766

Table 3: Pseudo pair accuracy and number of generated pairs, before and after applying ontology-aware filtering on the NCBI and BC5CDR training sets.

6.5.1 Cost-efficient Alias Generation for Lexical Diversity

Cost-efficiency and Reusability. Sci-ZSEL bounds LLM cost by querying aliases only for a selected subset of entities rather than every corpus mention. Comparing the ontology size (Table 1) against the selected set $\mathcal{E}_{\text{EM}} \cup \mathcal{E}_{\text{BT}}$ (Table 2), the efficiency gains are substantial for large ontologies: selected entities cover under 0.6% of MeSH and only 6.0% of MEDIC. For the smaller animal science ontologies, the percentages are unavoidably higher (31.9% for CMO, 35.0% for VT, 75.2% for LPT), a smaller subset would yield too few pseudo-labeled pairs for effective fine-tuning; however, the absolute number of processed entities remains modest (1,318, 1,416, and 391, respectively). Furthermore, since aliases are anchored to ontology entities, this supervision is reusable across any corpus linked to the same ontology, whereas per-mention augmentation is not reusable.

To further quantify computational efficiency, Table 6 compares Sci-ZSEL with per-mention augmentation in terms of the number of LLM calls and average input tokens per call. Sci-ZSEL requires fewer LLM calls on large ontologies and substantially fewer input tokens per call across all five benchmarks.

Dataset	Setting	BASE	RM-PCS	PC-POS
NCBI	No fine-tune	$88.12_{\pm 0.00}$	$88.12_{\pm 0.00}$	$88.12_{\pm 0.00}$
	Sci-ZSEL w/o filter	$87.74_{\pm 0.22}$	$87.01_{\pm 1.85}$	$\mathbf{88.75}_{\pm 0.36}$
	Sci-ZSEL	$87.40_{\pm 0.47}$	$80.70_{\pm 11.97}$	$\mathbf{87.81}_{\pm 0.63}$
	Synonym	$89.20_{\pm 0.24}$	$89.73_{\pm 0.30}$	$\mathbf{90.21}_{\pm 0.36}$
	Sci-ZSEL + Synonym	$89.72_{\pm 0.97}$	$89.41_{\pm 0.78}$	$\mathbf{90.80}_{\pm 0.87}$
BC5CDR	No fine-tune	$88.79_{\pm 0.00}$	$88.79_{\pm 0.00}$	$88.79_{\pm 0.00}$
	Sci-ZSEL w/o filter	$88.41_{\pm 0.74}$	$88.21_{\pm 2.01}$	$\mathbf{88.93}_{\pm 1.03}$
	Sci-ZSEL	$85.27_{\pm 1.75}$	$85.41_{\pm 0.27}$	$\mathbf{87.80}_{\pm 0.74}$
	Synonym	$87.22_{\pm 0.16}$	$87.13_{\pm 0.59}$	$\mathbf{88.57}_{\pm 0.50}$
	Sci-ZSEL + Synonym	$87.96_{\pm 0.48}$	$86.61_{\pm 2.16}$	$\mathbf{89.14}_{\pm 0.65}$
QTL$_{\text{CMO}}$	No fine-tune	$75.74_{\pm 0.00}$	$75.74_{\pm 0.00}$	$75.74_{\pm 0.00}$
	Sci-ZSEL w/o filter	$\mathbf{89.96}_{\pm 0.23}$	$88.02_{\pm 1.24}$	$89.24_{\pm 1.61}$
	Sci-ZSEL	$\mathbf{89.22}_{\pm 0.86}$	$88.06_{\pm 0.25}$	$88.34_{\pm 1.37}$
	Synonym	$\mathbf{85.71}_{\pm 0.40}$	$85.47_{\pm 0.49}$	$85.45_{\pm 0.91}$
	Sci-ZSEL + Synonym	$88.22_{\pm 0.89}$	$89.03_{\pm 1.68}$	$\mathbf{89.09}_{\pm 1.57}$
QTL$_{\text{VT}}$	No fine-tune	$72.63_{\pm 0.00}$	$72.63_{\pm 0.00}$	$72.63_{\pm 0.00}$
	Sci-ZSEL w/o filter	$\mathbf{84.70}_{\pm 0.96}$	$84.52_{\pm 1.08}$	$83.69_{\pm 0.94}$
	Sci-ZSEL	$84.72_{\pm 1.69}$	$\mathbf{85.51}_{\pm 0.30}$	$85.31_{\pm 1.03}$
	Synonym	$86.99_{\pm 1.74}$	$86.79_{\pm 2.67}$	$\mathbf{87.91}_{\pm 2.22}$
	Sci-ZSEL + Synonym	$85.58_{\pm 4.10}$	$\mathbf{88.69}_{\pm 2.22}$	$87.99_{\pm 1.47}$
QTL$_{\text{LPT}}$	No fine-tune	$68.70_{\pm 0.00}$	$68.70_{\pm 0.00}$	$68.70_{\pm 0.00}$
	Sci-ZSEL w/o filter	$70.08_{\pm 0.28}$	$70.82_{\pm 0.97}$	$\mathbf{71.01}_{\pm 0.92}$
	Sci-ZSEL	$75.07_{\pm 1.54}$	$\mathbf{76.87}_{\pm 0.73}$	$76.18_{\pm 0.37}$
	Synonym	$\mathbf{92.43}_{\pm 0.68}$	$91.83_{\pm 0.42}$	$91.78_{\pm 1.24}$
	Sci-ZSEL + Synonym	$92.29_{\pm 0.44}$	$92.48_{\pm 0.35}$	$\mathbf{92.66}_{\pm 0.69}$

Table 4: A comparison of retriever Recall@64 across benchmarks under various negative-sampling configurations. **Bold** indicates the best results when comparing column-wise.

Alias Lexical Diversity. Table 2 presents LLM-generated aliases by whether they match a curated synonym (%Syn), the entity name (%Name), or neither (constituting a new lexical bridge). On synonym-sparse animal science ontologies (CMO, VT, LPT), both %Syn and %Name are notably low. For instance, using the bi-encoder top-1 setting, over 80% of generated aliases serve as new lexical bridges, successfully providing diverse surface forms that the curated ontologies lack.

Pseudo-Pair Accuracy. LLM-based aliases are effective because they create lexical bridges, but they do not work reliably on their own because the raw alias drifts. However, our ontology-aware filter can remove drift. To evaluate the filter's effectiveness on our generated pseudo pairs, we compare against NCBI and BC5CDR, which provide gold labels for the training data. A pseudo-labeled sample is counted as correct whenever its assigned entity matches the gold label. Table 3 demonstrates that raw LLM alias generation (P_{Gen}`w/o filter`) includes noise, yielding a pseudo-label accuracy of 72% across both datasets. After removing drifted aliases (P_{Gen}), pseudo pair accuracy rises to over

Dataset	Setting	BLINK					ReS				
		Overall	MRR	HO	LO	NO	Overall	MRR	HO	LO	NO
NCBI	No fine-tune	64.38$_{\pm 0.00}$	72.39$_{\pm 0.00}$	**93.83**$_{\pm 0.00}$	64.12$_{\pm 0.00}$	23.64$_{\pm 0.00}$	45.94$_{\pm 0.00}$	58.41$_{\pm 0.00}$	67.53$_{\pm 0.00}$	44.44$_{\pm 0.00}$	18.64$_{\pm 0.00}$
	Sci-ZSEL w/o filter	70.69$_{\pm 0.58}$	77.65$_{\pm 0.47}$	92.75$_{\pm 0.19}$	66.44$_{\pm 1.22}$	48.18$_{\pm 0.79}$	60.28$_{\pm 0.42}$	73.57$_{\pm 0.27}$	68.72$_{\pm 0.68}$	59.41$_{\pm 1.88}$	50.15$_{\pm 3.19}$
	Sci-ZSEL	71.88$_{\pm 0.42}$	78.48$_{\pm 0.35}$	92.86$_{\pm 0.00}$	68.90$_{\pm 1.05}$	48.33$_{\pm 0.27}$	63.06$_{\pm 2.25}$	73.00$_{\pm 2.50}$	73.05$_{\pm 1.17}$	68.52$_{\pm 2.34}$	38.33$_{\pm 6.45}$
	Synonym	76.18$_{\pm 0.24}$	83.17$_{\pm 0.49}$	93.18$_{\pm 0.33}$	73.23$_{\pm 0.58}$	58.18$_{\pm 2.08}$	**78.09**$_{\pm 0.73}$	**84.59**$_{\pm 0.37}$	**93.94**$_{\pm 2.40}$	**74.07**$_{\pm 0.70}$	**63.79**$_{\pm 1.89}$
	Sci-ZSEL + Synonym	**77.43**$_{\pm 0.67}$	**84.35**$_{\pm 0.74}$	92.75$_{\pm 1.50}$	**74.54**$_{\pm 2.06}$	**61.67**$_{\pm 1.15}$	70.52$_{\pm 1.02}$	79.11$_{\pm 1.58}$	74.24$_{\pm 1.46}$	71.30$_{\pm 0.24}$	**63.79**$_{\pm 3.09}$
BC5CDR	No fine-tune	74.36$_{\pm 0.00}$	79.53$_{\pm 0.00}$	**96.26**$_{\pm 0.00}$	66.77$_{\pm 0.00}$	22.43$_{\pm 0.00}$	32.65$_{\pm 0.00}$	45.34$_{\pm 0.00}$	39.86$_{\pm 0.00}$	35.61$_{\pm 0.00}$	12.42$_{\pm 0.00}$
	Sci-ZSEL w/o filter	78.14$_{\pm 0.47}$	83.90$_{\pm 0.36}$	94.94$_{\pm 0.64}$	61.55$_{\pm 1.31}$	44.45$_{\pm 1.04}$	68.14$_{\pm 0.76}$	77.01$_{\pm 0.67}$	82.77$_{\pm 0.53}$	51.04$_{\pm 2.87}$	40.31$_{\pm 1.32}$
	Sci-ZSEL	79.16$_{\pm 0.58}$	84.40$_{\pm 0.27}$	96.01$_{\pm 0.31}$	72.17$_{\pm 2.03}$	39.88$_{\pm 1.45}$	76.91$_{\pm 0.53}$	82.52$_{\pm 0.28}$	**94.83**$_{\pm 0.58}$	64.65$_{\pm 1.58}$	37.85$_{\pm 1.42}$
	Synonym	78.20$_{\pm 1.24}$	83.49$_{\pm 1.88}$	96.17$_{\pm 0.24}$	67.97$_{\pm 1.93}$	37.85$_{\pm 3.51}$	74.74$_{\pm 0.40}$	80.15$_{\pm 0.48}$	94.12$_{\pm 0.44}$	58.68$_{\pm 3.04}$	34.09$_{\pm 0.83}$
	Sci-ZSEL + Synonym	**81.20**$_{\pm 0.58}$	**86.20**$_{\pm 0.48}$	95.96$_{\pm 0.30}$	**74.93**$_{\pm 1.96}$	**46.87**$_{\pm 1.30}$	**77.90**$_{\pm 0.12}$	**83.21**$_{\pm 0.26}$	94.78$_{\pm 0.17}$	**65.42**$_{\pm 2.49}$	**41.66**$_{\pm 0.66}$
QTL$_{\text{CMO}}$	No fine-tune	50.15$_{\pm 0.00}$	60.74$_{\pm 0.00}$	92.91$_{\pm 0.00}$	50.95$_{\pm 0.00}$	29.67$_{\pm 0.00}$	48.92$_{\pm 0.00}$	61.06$_{\pm 0.00}$	82.15$_{\pm 0.00}$	51.49$_{\pm 0.00}$	31.37$_{\pm 0.00}$
	Sci-ZSEL w/o filter	57.09$_{\pm 2.37}$	66.43$_{\pm 1.96}$	93.97$_{\pm 3.88}$	58.74$_{\pm 2.26}$	38.62$_{\pm 2.31}$	41.42$_{\pm 2.31}$	54.82$_{\pm 2.24}$	44.34$_{\pm 2.98}$	49.59$_{\pm 4.06}$	33.22$_{\pm 4.05}$
	Sci-ZSEL	56.89$_{\pm 0.48}$	67.15$_{\pm 0.34}$	93.16$_{\pm 0.42}$	56.89$_{\pm 0.14}$	**40.09**$_{\pm 1.08}$	56.84$_{\pm 1.46}$	66.72$_{\pm 1.46}$	96.74$_{\pm 0.79}$	58.43$_{\pm 1.97}$	**37.03**$_{\pm 2.98}$
	Synonym	57.59$_{\pm 1.72}$	66.50$_{\pm 1.51}$	93.15$_{\pm 0.25}$	62.79$_{\pm 2.24}$	36.77$_{\pm 2.34}$	**58.48**$_{\pm 0.37}$	**68.29**$_{\pm 0.36}$	94.70$_{\pm 0.51}$	64.91$_{\pm 0.41}$	36.32$_{\pm 1.09}$
	Sci-ZSEL + Synonym	**58.88**$_{\pm 3.30}$	**67.72**$_{\pm 2.92}$	**94.87**$_{\pm 0.65}$	**65.22**$_{\pm 2.99}$	36.88$_{\pm 5.58}$	57.43$_{\pm 0.23}$	66.95$_{\pm 0.12}$	**96.99**$_{\pm 0.79}$	**65.54**$_{\pm 0.36}$	32.31$_{\pm 0.91}$
QTL$_{\text{VT}}$	No fine-tune	43.13$_{\pm 0.00}$	56.38$_{\pm 0.00}$	97.32$_{\pm 0.00}$	48.70$_{\pm 0.00}$	31.43$_{\pm 0.00}$	38.33$_{\pm 0.00}$	51.14$_{\pm 0.00}$	95.97$_{\pm 0.00}$	54.61$_{\pm 0.00}$	19.71$_{\pm 0.00}$
	Sci-ZSEL w/o filter	38.29$_{\pm 1.21}$	54.51$_{\pm 0.74}$	97.32$_{\pm 0.00}$	50.90$_{\pm 2.58}$	21.64$_{\pm 0.94}$	45.81$_{\pm 2.13}$	59.78$_{\pm 0.86}$	97.32$_{\pm 0.00}$	64.12$_{\pm 5.84}$	26.94$_{\pm 0.40}$
	Sci-ZSEL	46.64$_{\pm 0.54}$	60.04$_{\pm 0.26}$	97.32$_{\pm 0.00}$	49.68$_{\pm 0.96}$	37.00$_{\pm 0.43}$	41.15$_{\pm 0.42}$	53.98$_{\pm 0.38}$	97.32$_{\pm 0.00}$	59.88$_{\pm 1.50}$	21.30$_{\pm 0.76}$
	Synonym	70.83$_{\pm 0.42}$	**79.66**$_{\pm 0.73}$	97.54$_{\pm 0.39}$	**76.64**$_{\pm 1.40}$	63.24$_{\pm 1.32}$	62.78$_{\pm 3.91}$	73.65$_{\pm 3.55}$	96.87$_{\pm 0.39}$	67.54$_{\pm 0.61}$	54.67$_{\pm 6.58}$
	Sci-ZSEL + Synonym	**71.25**$_{\pm 2.58}$	79.49$_{\pm 1.65}$	**97.77**$_{\pm 0.77}$	76.46$_{\pm 2.47}$	**64.04**$_{\pm 3.43}$	**65.58**$_{\pm 2.62}$	**75.76**$_{\pm 2.03}$	**97.77**$_{\pm 0.77}$	**74.09**$_{\pm 4.31}$	**55.53**$_{\pm 2.90}$
QTL$_{\text{LPT}}$	No fine-tune	57.34$_{\pm 0.00}$	63.29$_{\pm 0.00}$	94.59$_{\pm 0.00}$	78.29$_{\pm 0.00}$	20.62$_{\pm 0.00}$	59.70$_{\pm 0.00}$	67.91$_{\pm 0.00}$	92.34$_{\pm 0.00}$	77.71$_{\pm 0.00}$	27.69$_{\pm 0.00}$
	Sci-ZSEL w/o filter	54.62$_{\pm 1.06}$	61.67$_{\pm 0.85}$	95.05$_{\pm 0.46}$	72.00$_{\pm 2.29}$	17.64$_{\pm 1.45}$	55.82$_{\pm 0.64}$	64.20$_{\pm 0.84}$	93.54$_{\pm 2.03}$	68.95$_{\pm 1.44}$	22.97$_{\pm 0.89}$
	Sci-ZSEL	58.22$_{\pm 0.29}$	64.18$_{\pm 0.19}$	95.35$_{\pm 0.52}$	**79.81**$_{\pm 1.43}$	21.23$_{\pm 0.62}$	62.14$_{\pm 0.81}$	69.71$_{\pm 0.60}$	94.29$_{\pm 0.26}$	80.38$_{\pm 1.32}$	30.36$_{\pm 0.99}$
	Synonym	78.81$_{\pm 1.92}$	84.56$_{\pm 1.60}$	97.90$_{\pm 0.26}$	79.43$_{\pm 0.57}$	65.44$_{\pm 4.20}$	77.93$_{\pm 1.53}$	83.75$_{\pm 1.14}$	95.05$_{\pm 0.46}$	77.14$_{\pm 0.57}$	66.67$_{\pm 3.40}$
	Sci-ZSEL + Synonym	**80.70**$_{\pm 0.21}$	**86.12**$_{\pm 0.12}$	**98.05**$_{\pm 0.26}$	79.05$_{\pm 3.15}$	**69.74**$_{\pm 1.28}$	**81.21**$_{\pm 2.29}$	**86.21**$_{\pm 1.38}$	**97.45**$_{\pm 0.52}$	**80.95**$_{\pm 1.19}$	**70.26**$_{\pm 4.48}$

Table 5: A performance comparison of BLINK and ReS rerankers across all five benchmarks. The Recall@1 reported as "Overall" where the score considers the entire dataset, while HO, LO, and NO show category-wise scores. **Bold** indicates the best results when comparing row-wise.

	LLM calls		Input tokens	
Dataset	Ours	Per-ment.	Ours	Per-ment.
NCBI	**805**	4,584	**89.27**	302.83
BC5CDR	**2,062**	6,222	**77.85**	300.29
QTL$_{\text{CMO}}$	**1,318**	1,825	**66.07**	406.73
QTL$_{\text{VT}}$	**1,416**	2,134	**54.80**	388.59
QTL$_{\text{LPT}}$	**391**	694	**50.45**	395.98

Table 6: LLM call counts and average input tokens per call for entity-side selection (Ours, Sci-ZSEL) versus a per-mention augmentation baseline. Fewer is better; the lower value in each pair is in bold.

90% (90.32% for NCBI and 91.40% for BC5CDR). Although pseudo-pair accuracy cannot be directly measured for the animal-science benchmarks because gold training labels are unavailable, their downstream effect is reflected in the EL performance gains.

6.5.2 Retriever Results

Table 4 presents the retriever's performance, evaluated by overall Recall@64 under three negative-sampling strategies (BASE, RM-PCS, PC-POS), which we compare in Section 6.6.1; Here, we focus on the Sci-ZSEL + Synonym results under PC-POS. A key finding is that fine-tuning consistently improves results across all benchmarks, though improvements vary based on synonym sparsity. The magnitude of this improvement is heavily dependent on the dataset's inherent synonym density. On synonym-rich datasets like NCBI and BC5CDR, the non-fine-tuned backbone is already robust, resulting in relatively small fine-tuning gains (+2.68% and +0.35%, respectively). On synonym-sparse animal science datasets (QTL$_{\text{CMO}}$, QTL$_{\text{VT}}$, QTL$_{\text{LPT}}$), our ontology-aware fine-tuning delivers substantial improvements of +13.35% , +15.36% , and +23.96%, respectively.

6.5.3 Reranker Results

Table 5 presents the reranker results. Fine-tuning improves overall Recall@1 and MRR across every dataset and backbone architecture, where we observe the following major trends:

When utilizing the BLINK backbone, the combined Sci-ZSEL + Synonym approach is clearly the most robust, achieving the highest overall Recall@1 across all five benchmarks. The MRR results closely follow this trend, with Sci-ZSEL + Synonym dominating four of the five datasets. The single exception is the QTL$_{\text{VT}}$ dataset, where the standalone Synonym setting holds a marginal lead of +0.17% in MRR. For the ReS backbone, the results are more nuanced. Sci-ZSEL + Synonym remains the dominant setting for BC5CDR, QTL$_{\text{VT}}$, and QTL$_{\text{LPT}}$. However, on the NCBI and QTL$_{\text{CMO}}$ benchmarks, the Synonym setting performs best, outperforming the combined approach in Recall@1 by +7.57% and +1.05%, respectively. The MRR results for the ReS backbone mirror this exact trend, with the Synonym baseline maintaining its lead on those same two datasets.

Most notably, Sci-ZSEL excels in the highly

challenging NO cases across benchmarks, where curated lexical coverage is limited. Fine-tuning drives massive improvements in ZSEL, frequently doubling or tripling the non-fine-tuned backbone performance. Across all five benchmarks, we observe up to a +49.12%-point absolute gain in the NO category over the non-fine-tuned backbone across BLINK and ReS.

Collectively, these findings confirm that Sci-ZSEL effectively addresses a primary challenge of scientific EL. By utilizing LLM-generated aliases, the model successfully captures highly divergent surface forms, providing critical, complementary supervision that manually curated ontology synonyms alone cannot supply.

6.6 Ablation Studies

6.6.1 Retriever Ablations

We ablate two retriever design choices: the negative-sampling strategy and the pseudo-pair construction strategy.

For negative sampling, ontology-aware strategies outperform vanilla in-batch sampling, and PC-POS is the strongest overall, achieving the best Recall@64 on four of five benchmarks (Table 4); QTL_{VT} is the only exception.

For construction, the union of $\mathcal{P}_1$ and $\mathcal{P}_2$ (Sci-ZSEL) with synonyms is the strongest source of pseudo pairs on four of five benchmarks. Together these confirm that integrating ontology structure into both sampling and pseudo-pair construction improves retrieval. Full per-dataset results and discussion are in Appendix A.7.

6.6.2 Reranker Ablations

We ablate two reranker design choices: the ontology-aware filter and the pseudo-pair construction strategy.

The filter is necessary: removing it (Sci-ZSEL w/o filter) lets drifted aliases push reranker accuracy below the non-fine-tuned baseline on several dataset-backbone combinations (Table 5), most severely on the animal science benchmarks.

For construction, the pattern is backbone-dependent, with BLINK, the union (Sci-ZSEL) is best or tied-best on four of five benchmarks, whereas with ReS exact-match pairs alone are more competitive on NCBI and QTL_{CMO}, consistent with ReS being more sensitive to alias noise on these datasets. Full numbers are in Appendix A.7.

7 Acknowledgements

The work is supported in part by NSF-CAREER #2237831 and USDA-NIFA #2024-08585.

8 Conclusion

In conclusion, this paper offers a cost-aware solution to severe lexical divergence in scientific entity linking. We introduce Sci-ZSEL, a zero-shot framework that bounds LLM alias generation to selected entities and employs an ontology-aware filter to discard drifted aliases. To evaluate this, we also release a challenging, high-divergence animal science benchmark (QTL_{CMO}, QTL_{VT}, QTL_{LPT}). Across five datasets, Sci-ZSEL significantly outperforms non-fine-tuned baselines, achieving absolute gains in the lexically divergent domain. We demonstrate that combining Sci-ZSEL with curated synonyms yields the most robust reranker configuration, with the gain coming mainly from no-overlap mentions where curated synonyms alone fall short, and our ablation studies confirm the ontology-aware filter is essential to prevent LLM noise from degrading accuracy below baseline levels. Ultimately, this work provides a scalable, reusable blueprint for zero-shot domain adaptation in specialized, low-resource scientific domains.

9 Limitations

While Sci-ZSEL demonstrates strong performance in scientific entity linking, it is subject to several limitations. First, the framework relies heavily on the domain knowledge of the underlying LLM (Llama 3.2 3B Instruct in our experiments) to generate pseudo pairs. An LLM with weaker biomedical or animal science coverage would produce noisier, more drifted aliases, thereby increasing the burden on the filtering module. Additionally, future adaptations relying on closed-source LLMs could face reproducibility risks if silent updates arbitrarily alter alias distributions. Second, our current framework and evaluation are strictly limited to English. The benchmark datasets, prompt templates, filtering model (BioLORD), and reranking backbones (BLINK, ReS) are all predominantly English-centric. Extending this approach to other languages would require multilingual LLMs and semantic-similarity models specifically tailored to the biomedical domain, which are not currently available off-the-shelf.

Furthermore, the generated pseudo pairs and filtering decisions are intrinsically tied to fixed snapshots of the target ontologies (Appendix Table 9). Because ontologies are continuously updated with obsolete entries retired and hierarchical structures, such as the neighbor sets $\mathcal{N}(e)$ and $\mathcal{N}^{+}(e)$, frequently modified—any update to a target ontology requires Sci-ZSEL to regenerate the aliases and rerun the filtering process to avoid stale supervision. Finally, our current framework does not handle "NIL" entities, which are mentions lacking a valid referent in the target ontology. Although our new animal science benchmark reveals that a significant portion of mentions cannot be linked to all three ontologies, these NIL cases were excluded from our current evaluation. Developing a robust mechanism to identify and manage unlinkable mentions remains a critical practical challenge that we leave for future work.

10 Ethical considerations

Annotation process and compensation. The animal science benchmark (Section 3) was developed by a team comprising four graduate-student linking annotators and one domain-expert curator who was responsible for annotation validation and resolving disagreements. Annotation was performed by graduate-student members of the research team and curators of Animal QTLdb. The annotators were supported by NSF and USDA grants, and were not separately compensated. All source texts were derived entirely from publicly available PubMed articles, ensuring that the dataset contains no personally identifiable information (PII).

LLM usage disclosure. To generate entity aliases (Section 5.2), we utilized Llama 3.2 3B Instruct, an open-weights large language model. The LLM was prompted strictly using entity definitions sourced from public ontologies; at no point was human-generated or private data processed by the model. All other system components (BLINK, ReS, and BioLORD) are open-source and were deployed in strict adherence to their published usage licenses.

References

Alan R Aronson. 2001. Effective mapping of biomedical text to the UMLS metathesaurus: the MetaMap program. In *Proceedings of the AMIA Symposium*, pages 17–21.

Lihu Chen, Gaël Varoquaux, and Fabian M Suchanek. 2021. A lightweight neural model for biomedical entity linking. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 12657–12665.

Silviu Cucerzan. 2007. Large-scale named entity disambiguation based on wikipedia data. In *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*, pages 708–716.

Allan Peter Davis, Thomas C Wiegers, Michael C Rosenstein, and Carolyn J Mattingly. 2012. MEDIC: a practical disease vocabulary used at the comparative toxicogenomics database. *Database*, 2012:bar065.

Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. 2021. Autoregressive entity retrieval. In *International Conference on Learning Representations*.

Rezarta Islamaj Dogan, Robert Leaman, and Zhiyong Lu. 2014. NCBI disease corpus: a resource for disease name recognition and concept normalization. *Journal of biomedical informatics*, 47:1–10.

Dan Gillick, Sayali Kulkarni, Larry Lansing, Alessandro Presta, Jason Baldridge, Eugene Ie, and Diego Garcia-Olano. 2019. Learning dense representations for entity retrieval. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 528–537.

Zhi-Liang Hu, Eric Ryan Fritz, and James M. Reecy. 2007. AnimalQTLdb: a livestock QTL database tool set for positional QTL information mining and beyond. *Nucleic Acids Research*, 35(Database issue):D604–D609.

Chanhwi Kim, Hyunjae Kim, Sihyeon Park, Jiwoo Lee, Mujeen Sung, and Jaewoo Kang. 2025. Learning from negative samples in biomedical generative entity linking. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 10714–10730.

Robert Leaman, Rezarta Islamaj Dogan, and Zhiyong Lu. 2013. DNorm: disease name normalization with pairwise learning to rank. *Bioinformatics*, 29(22):2909–2917.

Jiao Li, Yueping Sun, Robin J Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Thomas C Wiegers, and Zhiyong Lu. 2016. BioCreative V CDR task corpus: a resource for chemical disease relation extraction. *Database*, 2016:baw068.

Yuepei Li, Kang Zhou, Qiao Qiao, Qing Wang, and Qi Li. 2025. Re-examine distantly supervised NER: A new benchmark and a simple approach. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10940–10959.

Carolyn E Lipscomb. 2000. Medical subject headings (MeSH). *Bulletin of the Medical Library Association*, 88(3):265–266.

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.

Lajanugen Logeswaran, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, Jacob Devlin, and Honglak Lee. 2019. Zero-shot entity linking by reading entity descriptions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3449–3460.

Lev Ratinov, Dan Roth, Doug Downey, and Mike Anderson. 2011. Local and global algorithms for disambiguation to wikipedia. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 1375–1384.

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*, 31(9):1844–1855.

Javier Sanz-Cruzado and Jake Lever. 2025. Accelerating cross-encoders in biomedical entity linking. In *Proceedings of the 24th Workshop on Biomedical Language Processing*, pages 136–147.

Mujeen Sung, Hwisang Jeon, Jinhyuk Lee, and Jaewoo Kang. 2020. Biomedical entity representations with synonym marginalization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3641–3650, Online. Association for Computational Linguistics.

Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. 2020. Scalable zero-shot entity linking with dense entity retrieval. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6397–6407.

Amy Xin, Yunjia Qi, Zijun Yao, Fangwei Zhu, Kaisheng Zeng, Bin Xu, Lei Hou, and Juanzi Li. 2025. LLMAEL: Large language models are good context augmenters for entity linking. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pages 3550–3559.

Zhenran Xu, Yulin Chen, and Baotian Hu. 2023a. Improving biomedical entity linking with cross-entity interaction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 13869–13877.

Zhenran Xu, Yulin Chen, Baotian Hu, and Min Zhang. 2023b. A read-and-select framework for zero-shot entity linking. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13657–13666.

Ikuya Yamada, Hiroyuki Shindo, Hideaki Takeda, and Yoshiyasu Takefuji. 2016. Joint learning of the embedding of words and entities for named entity disambiguation. In *Proceedings of the 20th SIGNLL conference on computational natural language learning*, pages 250–259.

Christophe Ye and Cassie S Mitchell. 2025. LLM as entity disambiguator for biomedical entity-linking. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 301–312.

Hongyi Yuan, Zheng Yuan, and Sheng Yu. 2022. Generative biomedical entity linking via knowledge base-guided pre-training and synonyms-aware fine-tuning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4038–4048.

Kang Zhou, Yuepei Li, Qing Wang, Qiao Qiao, and Qi Li. 2024. Gendecider: Integrating "none of the candidates" judgments in zero-shot entity linking reranking. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 239–245.

George Kingsley Zipf. 1949. *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*. Addison-Wesley, Cambridge, MA.

A Appendix

A.1 Entity Generation Prompt

We use a single prompt template across all five datasets, varying only the domain string and the three in-context examples, both of which are drawn from the target ontology. The template combines a system message specifying the domain expertise with a user message containing the definition to be named:

```
[System] You are a specialist in {domain} terminology. Given an entity's definition, generate a scientifically accurate name that best represents its meaning.

Guidelines:
1. Carefully interpret the semantic content of the "definition".
2. Generate a precise name that:
  - Accurately reflects the definition.
  - Aligns with standard terminology in {domain}.

Example 1
Definition: {def_1}
Generated name: {name_1}

Example 2
Definition: {def_2}
Generated name: {name_2}

Example 3
Definition: {def_3}
Generated name: {name_3}

[User] Definition : {definition}
Return all possible names for this definition, separated by commas. Do not include any additional text.
```

The domain string is *biomedical disease* for NCBI Disease, *biomedical Disease and biomedical Chemical* for BC5CDR, and *animal science* for QTL_{CMO}, QTL_{VT}, and QTL_{LPT}. The three in-context examples used for each ontology are listed in Table 14.

A.2 Number of Training Samples

Table 7 reports the number of pseudo-labeled pairs used for retriever and reranker fine-tuning under each setting. The number of samples (SMPL) vary by an order of magnitude across cells, driven by two factors: curated synonym density (Synonym yields 3,283 (NCBI), 4,676 (BC5CDR), 871 (QTL_{CMO}), 1,854 (QTL_{VT}), and 477 (QTL_{LPT}) pairs, tracking the synonym/entity ratios in Table 1 (ontology block)) and filter discard rate (Sci-ZSEL retains 74% (NCBI), 77% (BC5CDR), 64% (QTL_{CMO}), 29% (QTL_{VT}), and 22% (QTL_{LPT}) of unfiltered pairs, suggesting that the filter removes a larger fraction of generated aliases for the animal science ontologies. Sci-ZSEL +Synonym is the largest set across all benchmarks.

Setting	NCBI	BC5CDR	QTL_{CMO}	QTL_{VT}	QTL_{LPT}
Sci-ZSEL w/o filter	1757	7447	2513	1305	1592
Sci-ZSEL	1301	5766	1614	381	352
Synonym	3283	4676	871	1854	477
Sci-ZSEL + Synonym	4584	6222	1825	2134	694

Table 7: Number of training samples (SMPL) per (dataset, setting) used for fine-tuning.

A.3 Bi-encoder Details

A.3.1 Architecture

Following Wu et al. (2020), our bi-encoder uses two independent BERT transformers T_m and T_e to embed the mention context and the entity into a shared dense space. The mention input is

$$\tau_m = [\text{CLS}]\, \text{ctx}_l\, [\text{Ms}]\, m\, [\text{Me}]\, \text{ctx}_r\, [\text{SEP}]\,, \quad (11)$$

and the entity input is

$$\tau_e = [\text{CLS}]\, \text{title}_e\, [\text{ENT}]\, \text{desc}_e\, [\text{SEP}]\,, \quad (12)$$

where $\text{ctx}_l, \text{ctx}_r$ are the left and right contexts around m, and $\text{title}_e, \text{desc}_e$ are the entity title and description. Each sequence is encoded independently and the `[CLS]` output is taken as its dense representation:

$$y_m = \text{red}(T_m(\tau_m))\,, \quad (13)$$

$$y_e = \text{red}(T_e(\tau_e))\,. \quad (14)$$

The mention–entity score is the dot product

$$s(m, e) = y_m \cdot y_e\,. \quad (15)$$

The bi-encoder is initialized from the BLINK weights of Wu et al. (2020) and fine-tuned on our pseudo-labeled pairs.

A.3.2 Negative-sampling Formal Definitions

Notation. Let $B = \{m_1, \ldots, m_b\}$ be a training batch of b mentions with gold entities $\{e_1, \ldots, e_b\} \subset \mathcal{E}$. For each batch we construct a candidate column set $C = G \cup R$, where G is the set of unique golds in the batch and R is a set of random entities sampled uniformly from $\mathcal{E} \setminus G$.

Because the same gold entity may appear for multiple mentions in a batch, removing gold entities from the negative set can reduce the number of available in-batch negatives. We therefore add random entities R sampled from $\mathcal{E} \setminus G$ to keep the number of candidate negatives approximately stable. We reuse $\mathcal{N}(e)$ and $\mathcal{N}^+(e)$ from Section 5.2.2. For each row i with gold e_i, the positive set $\mathcal{P}_i$ and negative set $\mathcal{N}_i$ are defined per strategy in Table 8.

Strategy	Positive set $\mathcal{P}_i$	Negative set $\mathcal{N}_i$
BASE	$\{e_i\}$	$(G \setminus \{e_i\}) \cup R$
RM-PCS	$\{e_i\}$	$\big((G \setminus \{e_i\}) \cup R\big) \setminus \mathcal{N}^+(e_i)$
PC-POS	$\{e_i\} \cup \big(\mathcal{N}(e_i) \cap G\big)$	$\big((G \setminus \mathcal{P}_i) \cup R\big) \setminus \mathcal{N}(e_i)$

Table 8: Per-row positive set $\mathcal{P}_i$ and negative set $\mathcal{N}_i$ for the three negative-sampling strategies. $\mathcal{N}(e_i) =$ parents$(e_i) \cup$ children(e_i); $\mathcal{N}^+(e_i) = \mathcal{N}(e_i) \cup$ siblings(e_i). Siblings are not promoted as positives in PC-POS, preserving sibling separation pressure.

Sampling budget. Naively sampling $|R| = b - |G|$ leaves $\mathcal{N}_i$ short whenever RM-PCS or PC-POS discards neighbors. To keep the per-row negative count stable, we sample $|R| = (b - |G|) + \min(\max_i |\mathcal{N}^+(e_i)|, H_{\text{cap}})$ with $H_{\text{cap}} = 30$, then truncate each row's $\mathcal{N}_i$ to the target budget.

A.3.3 Training Objective

Following the in-batch cross-entropy objective of Wu et al. (2020), for each row i the loss is

$$\mathcal{L}_i = -\frac{1}{|\mathcal{P}_i|} \sum_{p \in \mathcal{P}_i} \log \frac{\exp s(m_i, p)}{Z_i(p)} , \qquad (16)$$

where the partition function is

$$Z_i(p) = \exp s(m_i, p) + \sum_{c \in \mathcal{N}_i} \exp s(m_i, c) .$$

For BASE and RM-PCS, $|\mathcal{P}_i| = 1$ and the expression reduces to the standard BLINK in-batch cross-entropy loss; for PC-POS, the average is taken over multiple positives against the same negative set. The full batch loss is

$$\mathcal{L} = \frac{1}{b} \sum_{i=1}^{b} \mathcal{L}_i . \qquad (17)$$

A.4 Cross-encoder Details

A.4.1 Architecture

We evaluate two cross-encoder backbones, BLINK (Wu et al., 2020) and ReS (Xu et al., 2023b). Both share the same concatenated mention–entity input:

$$\begin{aligned} \tau_{m,e} = &[\text{CLS}]\ \text{ctx}_l\ [\text{Ms}]\ m\ [\text{Me}]\ \text{ctx}_r\ [\text{SEP}] \\ &\text{title}_e\ [\text{ENT}]\ \text{desc}_e\ [\text{SEP}] , \end{aligned} \qquad (18)$$

where $\text{ctx}_l, \text{ctx}_r$ are the left and right contexts around m, and $\text{title}_e, \text{desc}_e$ are the entity title and description.

BLINK cross-encoder. Following Wu et al. (2020), a single BERT transformer T_{cross} jointly encodes $\tau_{m,e}$, and the score is a linear projection of the reduced [CLS] embedding:

$$y_{m,e} = \text{red}(T_{\text{cross}}(\tau_{m,e})) , \qquad (19)$$

$$s_{\text{BLINK}}(m, e) = \mathbf{w}^\top y_{m,e} , \qquad (20)$$

where $\mathbf{w}$ is a learned scoring vector. The BLINK reranker is initialized from the cross-encoder weights of Wu et al. (2020) pretrained on Wikipedia entity linking.

ReS cross-encoder. The Read-and-Select framework (Xu et al., 2023b) factors scoring into two stages. In the *reading* stage, each candidate $e \in \mathcal{C}(m)$ is encoded with the mention by a cross-encoder T_{read} to produce a candidate-conditioned mention representation h_m^e. In the *selecting* stage, the $|\mathcal{C}(m)|$ representations are passed through a candidate-level transformer that performs cross-candidate attention, so the score for e depends on the other candidates in $\mathcal{C}(m)$ rather than on e alone:

$$s_{\text{ReS}}(m, e) = f_{\text{sel}}\big(\{h_m^{e'}\}_{e' \in \mathcal{C}(m)},\, e\big) , \qquad (21)$$

where f_{sel} is the selecting head; see Xu et al. (2023b) for the full architecture. The ReS reranker is initialized from the released ReS weights.

A.4.2 Candidate Set Construction

For each pseudo pair $\langle m, e \rangle$, let $\{c_1, \ldots, c_K\}$ be the top-K candidates from the BLINK bi-encoder, ranked by retriever score. The reranker candidate set $\mathcal{C}(m)$ equals $\{c_1, \ldots, c_K\}$ when $e \in \{c_1, \ldots, c_K\}$, and otherwise replaces the lowest-ranked candidate with e, giving $\{c_1, \ldots, c_{K-1}, e\}$. This guarantees $e \in \mathcal{C}(m)$ and $|\mathcal{C}(m)| = K$.

We set K=64. After ensuring that the pseudo-labeled entity e is included in the candidate set, e is treated as the single positive and the remaining 63 candidates as negatives. For reranker fine-tuning, we randomly sample 20 negative examples from these 63 negatives and train the reranker using the

positive entity together with the 20 sampled negatives. The reranker is then trained on the reduced candidate set whereas the full candidate set is retained at inference, where the reranker scores all K retrieved candidates.

A.4.3 Training Objective

For each pseudo pair $\langle m, e \rangle$ with candidate set $\mathcal{C}(m)$, the reranker is trained with cross-entropy, treating e as the positive and the remaining candidates as negatives:

$$\mathcal{L} = -\log \frac{\exp s(m,e)}{\sum_{c \in \mathcal{C}(m)} \exp s(m,c)}, \qquad (22)$$

where $s(m,c)$ is the reranker score (s_{BLINK} or s_{ReS}). The same objective is used for both backbones; only the scoring function differs.

A.5 Training Hyperparameters and Hardware

Hardware. LLM alias generation uses Llama 3.2 3B Instruct on a single NVIDIA A100. Retriever and reranker fine-tuning use four A100s.

Bi-encoder retriever. The BLINK retriever uses `bert-large-uncased` with a 128-token mention context and 128-token candidate. It is trained at a learning rate of 2×10^{-5} with batch size 512, falling back to 128 when fewer than 512 training pairs are available.

Cross-encoder rerankers. The BLINK reranker uses `bert-large-uncased` with a 64-token context and 128-token candidate, trained at learning rate 2×10^{-5} with batch size 32. The ReS reranker uses `roberta-base` with a 256-token context and 256-token entity description, trained at learning rate 1×10^{-4} with batch size 8. Dropout is fixed at 0.2 throughout.

Epochs and reporting protocol. Because the zero-shot setting provides no validation set, best-epoch selection is not possible; we fix the number of epochs in advance and report results from the last epoch. The retriever runs for 4 epochs on BC5CDR (the largest dataset) and 1 epoch on the remaining four datasets. The reranker runs for 3 epochs on all benchmarks. Empirical loss curves confirm that the chosen epoch counts place the final epoch at or near the training optimum for nearly all benchmark–backbone–seed combinations,[2] validating our last-epoch reporting protocol.

A.6 Ontology/Knowledge Base Versions

Table 9 reports the specific releases of each knowledge base used in our experiments. Versions were fixed at the start of experimentation and reused across all settings.

KB / Ontology	Used by	Release
MeSH 2025	BC5CDR	2025-01-01
MEDIC	NCBI Disease	2025-02-28
CMO	QTL_{CMO}	2026-02-28
VT	QTL_{VT}	2026-03-03
LPT	QTL_{LPT}	2025-08-14

Table 9: Release versions of the knowledge bases and ontologies used in this work.

A.7 Detailed Ablation Analysis

Retriever: negative sampling. We compare BASE, RM-PCS, and PC-POS under the Sci-ZSEL + Synonym framework (Table 4). PC-POS achieves the highest Recall@64 on NCBI, BC5CDR, QTL_{CMO}, and QTL_{LPT}. The sole exception is QTL_{VT}, where RM-PCS peaks at 88.69%; for VT's ontological structure, removing sibling entities from the negatives (RM-PCS) yields a stronger signal than adding neighbor entities as positives (PC-POS). The MRR results (Tables 10, 11, 12) follow a broadly similar trend: PC-POS achieves the highest MRR on NCBI, BC5CDR, and QTL_{LPT}, RM-PCS on QTL_{VT}, and BASE on QTL_{CMO}. These results confirm that integrating ontology structure into sampling is superior to vanilla construction.

Retriever: construction strategy. We ablate the pseudo-pair construction strategy in Table 13. Sci-ZSEL + Synonym is the best strategy on NCBI, BC5CDR, QTL_{CMO}, and QTL_{LPT}, showing that the union of $\mathcal{P}_1$ (Eq. 3) and $\mathcal{P}_2$ (Eq. 8) with synonyms is the strongest source of pseudo pairs. QTL_{VT} is the only exception, where $\mathcal{P}_2$ + Synonym leads by a small margin (89.18% vs. 87.99%). Both construction strategies contribute to retrieval, and combining them is generally preferable.

[2] The only exception is the ReS reranker on BC5CDR under $\mathcal{P}_1+$ Synonym. One of the three seeds becomes unstable after epoch 1: the training loss increases, while test accuracy drops from about 75% to 10%. Following our fixed-epoch protocol, we report this run without modification. This unstable seed leads to the high variance (50.98 ± 41.58) reported for this setting in Table 13, while the other two seeds remain stable.

Reranker: ontology-aware filter. We verify the necessity of the filter by omitting it (Sci-ZSEL w/o filter). Removing the filter often harms performance, causing reranker accuracy (Table 5) to drop below the non-fine-tuned baseline across several dataset–backbone combinations. The largest drop in overall Recall@1 is 7.50% on QTL_{CMO} with ReS. This occurs because drifted aliases train the model to link mentions to a neighboring entity rather than the true referent; as a result, performance drops even on the easiest HO cases (the largest HO Recall@1 drop is 37.81%). These effects are most severe on the animal science datasets. We conclude that the ontology-aware filter is necessary to leverage LLM-based aliases, especially in low-resource settings.

Reranker: construction strategy. We ablate construction under both backbones in Table 13. With BLINK, Sci-ZSEL + Synonym is best or tied-best on four of five datasets, again favoring the union of $\mathcal{P}_1$ and $\mathcal{P}_2$. With ReS, exact-match alone ($\mathcal{P}_1$ + Synonym) is more competitive on NCBI and QTL_{CMO}, and on these two datasets ReS also performs best using ontology synonyms alone (Table 5). This suggests ReS is more sensitive to training noise on these datasets, so the added LLM aliases help less there.

Dataset	Setting	Overall	MRR	HO	LO	NO
NCBI	No fine-tune	88.12$_{\pm 0.00}$	61.94$_{\pm 0.00}$	**99.68**$_{\pm 0.00}$	92.82$_{\pm 0.00}$	62.73$_{\pm 0.00}$
	Sci-ZSEL w/o filter	87.74$_{\pm 0.22}$	62.48$_{\pm 1.82}$	99.46$_{\pm 0.19}$	91.67$_{\pm 0.23}$	63.64$_{\pm 1.58}$
	Sci-ZSEL	87.40$_{\pm 0.47}$	58.11$_{\pm 2.28}$	**99.68**$_{\pm 0.00}$	90.97$_{\pm 0.61}$	63.18$_{\pm 1.21}$
	Synonym	89.20$_{\pm 0.24}$	**64.53**$_{\pm 2.28}$	99.46$_{\pm 0.38}$	92.05$_{\pm 0.35}$	69.24$_{\pm 0.95}$
	Sci-ZSEL + Synonym	**89.72**$_{\pm 0.97}$	64.35$_{\pm 2.10}$	**99.68**$_{\pm 0.00}$	**92.90**$_{\pm 1.33}$	**69.54**$_{\pm 1.64}$
BC5CDR	No fine-tune	**88.79**$_{\pm 0.00}$	73.91$_{\pm 0.00}$	99.46$_{\pm 0.00}$	**96.09**$_{\pm 0.00}$	**57.24**$_{\pm 0.00}$
	Sci-ZSEL w/o filter	88.41$_{\pm 0.74}$	**74.09**$_{\pm 1.22}$	**99.49**$_{\pm 0.07}$	95.52$_{\pm 1.05}$	55.88$_{\pm 2.40}$
	Sci-ZSEL	85.27$_{\pm 1.75}$	70.75$_{\pm 2.08}$	99.23$_{\pm 0.23}$	89.20$_{\pm 3.75}$	47.17$_{\pm 4.53}$
	Synonym	87.22$_{\pm 0.16}$	72.03$_{\pm 0.26}$	99.25$_{\pm 0.10}$	91.64$_{\pm 1.08}$	53.78$_{\pm 0.45}$
	Sci-ZSEL + Synonym	87.96$_{\pm 0.48}$	73.20$_{\pm 0.33}$	99.40$_{\pm 0.04}$	93.55$_{\pm 1.01}$	55.39$_{\pm 1.60}$
QTL$_{\text{CMO}}$	No fine-tune	75.74$_{\pm 0.00}$	43.20$_{\pm 0.00}$	99.27$_{\pm 0.00}$	87.43$_{\pm 0.00}$	55.04$_{\pm 0.00}$
	Sci-ZSEL w/o filter	**89.96**$_{\pm 0.23}$	54.09$_{\pm 2.73}$	**99.51**$_{\pm 0.00}$	97.48$_{\pm 0.08}$	**79.24**$_{\pm 0.47}$
	Sci-ZSEL	89.22$_{\pm 0.86}$	56.06$_{\pm 0.55}$	**99.51**$_{\pm 0.00}$	97.97$_{\pm 0.24}$	77.12$_{\pm 2.16}$
	Synonym	85.71$_{\pm 0.40}$	53.66$_{\pm 0.89}$	99.35$_{\pm 0.14}$	97.12$_{\pm 0.39}$	69.84$_{\pm 0.54}$
	Sci-ZSEL + Synonym	88.22$_{\pm 0.89}$	**58.17**$_{\pm 1.01}$	**99.51**$_{\pm 0.00}$	**98.29**$_{\pm 0.20}$	74.56$_{\pm 1.93}$
QTL$_{\text{VT}}$	No fine-tune	72.63$_{\pm 0.00}$	34.67$_{\pm 0.00}$	**99.33**$_{\pm 0.00}$	86.96$_{\pm 0.00}$	59.96$_{\pm 0.00}$
	Sci-ZSEL w/o filter	84.70$_{\pm 0.96}$	43.95$_{\pm 1.20}$	**99.33**$_{\pm 0.00}$	96.64$_{\pm 0.44}$	75.31$_{\pm 1.53}$
	Sci-ZSEL	84.72$_{\pm 1.69}$	45.20$_{\pm 0.90}$	**99.33**$_{\pm 0.00}$	**96.75**$_{\pm 0.40}$	75.28$_{\pm 2.73}$
	Synonym	**86.99**$_{\pm 1.74}$	**47.83**$_{\pm 4.78}$	**99.33**$_{\pm 0.00}$	93.22$_{\pm 2.30}$	**81.36**$_{\pm 1.82}$
	Sci-ZSEL + Synonym	85.58$_{\pm 4.10}$	45.50$_{\pm 15.06}$	98.43$_{\pm 1.55}$	91.88$_{\pm 4.37}$	79.84$_{\pm 4.55}$
QTL$_{\text{LPT}}$	No fine-tune	68.70$_{\pm 0.00}$	51.17$_{\pm 0.00}$	**100.00**$_{\pm 0.00}$	90.86$_{\pm 0.00}$	35.38$_{\pm 0.00}$
	Sci-ZSEL w/o filter	70.08$_{\pm 0.28}$	47.50$_{\pm 0.96}$	**100.00**$_{\pm 0.00}$	93.33$_{\pm 0.33}$	37.13$_{\pm 0.47}$
	Sci-ZSEL	75.07$_{\pm 1.54}$	54.76$_{\pm 0.14}$	**100.00**$_{\pm 0.00}$	**96.19**$_{\pm 0.33}$	46.67$_{\pm 3.46}$
	Synonym	**92.43**$_{\pm 0.68}$	**75.56**$_{\pm 0.87}$	**100.00**$_{\pm 0.00}$	91.05$_{\pm 2.01}$	**88.00**$_{\pm 0.62}$
	Sci-ZSEL + Synonym	92.29$_{\pm 0.44}$	57.38$_{\pm 1.37}$	**100.00**$_{\pm 0.00}$	92.19$_{\pm 0.66}$	87.08$_{\pm 0.93}$

Table 10: Retriever Recall@64 under the **BASE** negative-sampling strategy. **Bold** represents the best results when comparing row-wise.

Dataset	Setting	Overall	MRR	HO	LO	NO
NCBI	No fine-tune	88.12$_{\pm 0.00}$	61.94$_{\pm 0.00}$	**99.68**$_{\pm 0.00}$	**92.82**$_{\pm 0.00}$	62.73$_{\pm 0.00}$
	Sci-ZSEL w/o filter	87.01$_{\pm 1.85}$	63.75$_{\pm 1.48}$	98.70$_{\pm 1.42}$	90.67$_{\pm 1.74}$	63.48$_{\pm 2.96}$
	Sci-ZSEL	80.70$_{\pm 11.97}$	54.37$_{\pm 10.62}$	92.21$_{\pm 12.94}$	83.26$_{\pm 13.97}$	59.55$_{\pm 6.70}$
	Synonym	**89.73**$_{\pm 0.30}$	63.99$_{\pm 1.16}$	**99.68**$_{\pm 0.00}$	92.36$_{\pm 1.16}$	**70.60**$_{\pm 1.31}$
	Sci-ZSEL + Synonym	89.41$_{\pm 0.78}$	**64.41**$_{\pm 1.94}$	**99.68**$_{\pm 0.00}$	91.66$_{\pm 0.80}$	**70.60**$_{\pm 2.15}$
BC5CDR	No fine-tune	**88.79**$_{\pm 0.00}$	**73.91**$_{\pm 0.00}$	**99.46**$_{\pm 0.00}$	**96.09**$_{\pm 0.00}$	**57.24**$_{\pm 0.00}$
	Sci-ZSEL w/o filter	88.21$_{\pm 2.01}$	73.51$_{\pm 2.57}$	99.42$_{\pm 0.23}$	95.39$_{\pm 3.00}$	55.30$_{\pm 6.11}$
	Sci-ZSEL	85.41$_{\pm 0.27}$	70.29$_{\pm 0.66}$	99.29$_{\pm 0.15}$	89.20$_{\pm 1.98}$	47.63$_{\pm 2.14}$
	Synonym	87.13$_{\pm 0.59}$	72.24$_{\pm 0.58}$	99.34$_{\pm 0.19}$	91.87$_{\pm 0.86}$	53.08$_{\pm 2.85}$
	Sci-ZSEL + Synonym	86.61$_{\pm 2.16}$	68.05$_{\pm 7.57}$	99.00$_{\pm 0.56}$	90.18$_{\pm 5.00}$	52.75$_{\pm 4.63}$
QTL$_{\text{CMO}}$	No fine-tune	75.74$_{\pm 0.00}$	43.20$_{\pm 0.00}$	99.27$_{\pm 0.00}$	87.43$_{\pm 0.00}$	55.04$_{\pm 0.00}$
	Sci-ZSEL w/o filter	88.02$_{\pm 1.24}$	54.37$_{\pm 1.05}$	**99.51**$_{\pm 0.00}$	95.68$_{\pm 2.39}$	76.29$_{\pm 0.86}$
	Sci-ZSEL	88.06$_{\pm 0.25}$	56.00$_{\pm 0.95}$	**99.51**$_{\pm 0.00}$	97.97$_{\pm 0.24}$	74.44$_{\pm 0.51}$
	Synonym	85.47$_{\pm 0.49}$	53.47$_{\pm 0.35}$	99.27$_{\pm 0.00}$	97.07$_{\pm 0.16}$	69.35$_{\pm 1.05}$
	Sci-ZSEL + Synonym	**89.03**$_{\pm 1.68}$	**57.17**$_{\pm 0.99}$	**99.51**$_{\pm 0.00}$	**98.33**$_{\pm 0.34}$	**76.37**$_{\pm 3.58}$
QTL$_{\text{VT}}$	No fine-tune	72.63$_{\pm 0.00}$	34.67$_{\pm 0.00}$	**99.33**$_{\pm 0.00}$	86.96$_{\pm 0.00}$	59.96$_{\pm 0.00}$
	Sci-ZSEL w/o filter	84.52$_{\pm 1.08}$	44.05$_{\pm 0.51}$	**99.33**$_{\pm 0.00}$	96.41$_{\pm 0.36}$	75.14$_{\pm 1.79}$
	Sci-ZSEL	85.51$_{\pm 0.30}$	43.49$_{\pm 0.72}$	**99.33**$_{\pm 0.00}$	**96.64**$_{\pm 0.44}$	76.73$_{\pm 0.26}$
	Synonym	86.79$_{\pm 2.67}$	46.21$_{\pm 3.65}$	**99.33**$_{\pm 0.00}$	93.04$_{\pm 3.31}$	81.12$_{\pm 2.81}$
	Sci-ZSEL + Synonym	**88.69**$_{\pm 2.22}$	**49.50**$_{\pm 4.52}$	**99.33**$_{\pm 0.00}$	94.32$_{\pm 3.26}$	**83.68**$_{\pm 1.95}$
QTL$_{\text{LPT}}$	No fine-tune	68.70$_{\pm 0.00}$	51.17$_{\pm 0.00}$	**100.00**$_{\pm 0.00}$	90.86$_{\pm 0.00}$	35.38$_{\pm 0.00}$
	Sci-ZSEL w/o filter	70.82$_{\pm 0.97}$	48.46$_{\pm 1.21}$	**100.00**$_{\pm 0.00}$	93.71$_{\pm 0.58}$	38.57$_{\pm 1.85}$
	Sci-ZSEL	76.87$_{\pm 0.73}$	54.03$_{\pm 1.07}$	**100.00**$_{\pm 0.00}$	**95.24**$_{\pm 1.19}$	51.18$_{\pm 2.00}$
	Synonym	91.83$_{\pm 0.42}$	**75.45**$_{\pm 1.19}$	**100.00**$_{\pm 0.00}$	90.10$_{\pm 1.84}$	87.18$_{\pm 0.35}$
	Sci-ZSEL + Synonym	**92.48**$_{\pm 0.35}$	57.37$_{\pm 0.59}$	**100.00**$_{\pm 0.00}$	92.57$_{\pm 1.14}$	**87.28**$_{\pm 0.17}$

Table 11: Retriever Recall@64 under the **RM-PCS** negative-sampling strategy. **Bold** represents the best results when comparing row-wise.

Dataset	Setting	Overall	MRR	HO	LO	NO
NCBI	No fine-tune	88.12$_{\pm 0.00}$	61.94$_{\pm 0.00}$	**99.68**$_{\pm 0.00}$	**92.82**$_{\pm 0.00}$	62.73$_{\pm 0.00}$
	Sci-ZSEL w/o filter	88.75$_{\pm 0.36}$	63.78$_{\pm 1.61}$	99.46$_{\pm 0.19}$	92.21$_{\pm 0.53}$	66.97$_{\pm 2.89}$
	Sci-ZSEL	87.81$_{\pm 0.63}$	61.12$_{\pm 1.48}$	**99.68**$_{\pm 0.00}$	91.82$_{\pm 0.93}$	63.33$_{\pm 0.95}$
	Synonym	90.21$_{\pm 0.36}$	65.20$_{\pm 0.97}$	99.35$_{\pm 0.33}$	91.98$_{\pm 0.48}$	73.94$_{\pm 0.26}$
	Sci-ZSEL + Synonym	**90.80**$_{\pm 0.87}$	**66.63**$_{\pm 1.46}$	99.46$_{\pm 0.38}$	91.98$_{\pm 0.48}$	**76.37**$_{\pm 2.36}$
BC5CDR	No fine-tune	88.79$_{\pm 0.00}$	73.91$_{\pm 0.00}$	99.46$_{\pm 0.00}$	96.09$_{\pm 0.00}$	57.24$_{\pm 0.00}$
	Sci-ZSEL w/o filter	88.93$_{\pm 1.03}$	73.76$_{\pm 1.74}$	**99.53**$_{\pm 0.04}$	**96.19**$_{\pm 1.62}$	57.54$_{\pm 3.34}$
	Sci-ZSEL	87.80$_{\pm 0.74}$	72.63$_{\pm 0.97}$	99.42$_{\pm 0.10}$	94.60$_{\pm 1.50}$	54.09$_{\pm 2.09}$
	Synonym	88.57$_{\pm 0.50}$	73.50$_{\pm 0.51}$	99.48$_{\pm 0.05}$	95.50$_{\pm 0.62}$	56.58$_{\pm 2.24}$
	Sci-ZSEL + Synonym	**89.14**$_{\pm 0.65}$	**74.05**$_{\pm 0.46}$	99.29$_{\pm 0.20}$	95.60$_{\pm 1.58}$	**59.38**$_{\pm 1.69}$
QTL$_{\text{CMO}}$	No fine-tune	75.74$_{\pm 0.00}$	43.20$_{\pm 0.00}$	99.27$_{\pm 0.00}$	87.43$_{\pm 0.00}$	55.04$_{\pm 0.00}$
	Sci-ZSEL w/o filter	**89.24**$_{\pm 1.61}$	54.53$_{\pm 1.76}$	**99.51**$_{\pm 0.00}$	96.89$_{\pm 1.18}$	**78.07**$_{\pm 2.77}$
	Sci-ZSEL	88.34$_{\pm 1.37}$	54.79$_{\pm 1.08}$	**99.51**$_{\pm 0.00}$	97.79$_{\pm 0.16}$	75.23$_{\pm 3.14}$
	Synonym	85.45$_{\pm 0.91}$	53.31$_{\pm 0.96}$	99.43$_{\pm 0.14}$	97.16$_{\pm 0.41}$	69.16$_{\pm 1.71}$
	Sci-ZSEL + Synonym	89.09$_{\pm 1.57}$	**56.33**$_{\pm 0.35}$	**99.51**$_{\pm 0.00}$	**98.33**$_{\pm 0.20}$	76.52$_{\pm 3.46}$
QTL$_{\text{VT}}$	No fine-tune	72.63$_{\pm 0.00}$	34.67$_{\pm 0.00}$	**99.33**$_{\pm 0.00}$	86.96$_{\pm 0.00}$	59.96$_{\pm 0.00}$
	Sci-ZSEL w/o filter	83.69$_{\pm 0.94}$	43.61$_{\pm 0.66}$	**99.33**$_{\pm 0.00}$	96.23$_{\pm 0.36}$	73.79$_{\pm 1.51}$
	Sci-ZSEL	85.31$_{\pm 1.03}$	45.32$_{\pm 0.93}$	**99.33**$_{\pm 0.00}$	**96.70**$_{\pm 0.30}$	76.35$_{\pm 1.67}$
	Synonym	87.91$_{\pm 2.22}$	46.94$_{\pm 2.40}$	**99.33**$_{\pm 0.00}$	93.33$_{\pm 1.86}$	82.92$_{\pm 2.94}$
	Sci-ZSEL + Synonym	**87.99**$_{\pm 1.47}$	**47.63**$_{\pm 2.73}$	**99.33**$_{\pm 0.00}$	92.69$_{\pm 3.08}$	**83.44**$_{\pm 1.24}$
QTL$_{\text{LPT}}$	No fine-tune	68.70$_{\pm 0.00}$	51.17$_{\pm 0.00}$	**100.00**$_{\pm 0.00}$	90.86$_{\pm 0.00}$	35.38$_{\pm 0.00}$
	Sci-ZSEL w/o filter	71.01$_{\pm 0.92}$	50.47$_{\pm 1.78}$	**100.00**$_{\pm 0.00}$	94.86$_{\pm 1.51}$	38.36$_{\pm 1.24}$
	Sci-ZSEL	76.18$_{\pm 0.37}$	54.01$_{\pm 0.29}$	**100.00**$_{\pm 0.00}$	**95.62**$_{\pm 1.19}$	49.44$_{\pm 1.24}$
	Synonym	91.78$_{\pm 1.24}$	**75.72**$_{\pm 0.62}$	**100.00**$_{\pm 0.00}$	90.28$_{\pm 2.49}$	86.97$_{\pm 1.43}$
	Sci-ZSEL + Synonym	**92.66**$_{\pm 0.69}$	57.61$_{\pm 0.65}$	**100.00**$_{\pm 0.00}$	93.14$_{\pm 1.15}$	**87.38**$_{\pm 0.93}$

Table 12: Retriever Recall@64 under the **PC-POS** negative-sampling strategy. **Bold** represents the best results when comparing row-wise.

Dataset	Setting	Retriever	BLINK reranker		ReS reranker	
		Recall@64	Recall@1	MRR	Recall@1	MRR
NCBI	$\mathcal{P}_1$ + Synonym	90.56$_{\pm 0.16}$	77.15$_{\pm 0.73}$	83.88$_{\pm 0.36}$	**78.26**$_{\pm 0.91}$	**84.46**$_{\pm 0.79}$
	$\mathcal{P}_2$ + Synonym	90.42$_{\pm 0.68}$	74.93$_{\pm 4.49}$	82.95$_{\pm 2.22}$	70.69$_{\pm 0.57}$	80.06$_{\pm 1.07}$
	Sci-ZSEL + Synonym	**90.80**$_{\pm 0.87}$	**77.43**$_{\pm 0.67}$	**84.35**$_{\pm 0.74}$	70.52$_{\pm 1.02}$	79.11$_{\pm 1.58}$
BC5CDR	$\mathcal{P}_1$ + Synonym	88.47$_{\pm 0.35}$	79.35$_{\pm 0.38}$	84.98$_{\pm 0.42}$	50.98$_{\pm 41.58}$	56.75$_{\pm 40.31}$
	$\mathcal{P}_2$ + Synonym	88.71$_{\pm 0.83}$	81.19$_{\pm 0.59}$	**86.33**$_{\pm 0.65}$	77.37$_{\pm 0.67}$	82.91$_{\pm 0.62}$
	Sci-ZSEL + Synonym	**89.14**$_{\pm 0.65}$	**81.20**$_{\pm 0.58}$	86.20$_{\pm 0.48}$	**77.90**$_{\pm 0.12}$	**83.21**$_{\pm 0.26}$
QTL$_{\text{CMO}}$	$\mathcal{P}_1$ + Synonym	87.45$_{\pm 0.50}$	57.63$_{\pm 2.02}$	66.97$_{\pm 1.85}$	**58.60**$_{\pm 1.94}$	**68.09**$_{\pm 1.63}$
	$\mathcal{P}_2$ + Synonym	87.01$_{\pm 0.60}$	**59.28**$_{\pm 0.36}$	**68.27**$_{\pm 0.44}$	57.67$_{\pm 0.92}$	67.82$_{\pm 0.81}$
	Sci-ZSEL + Synonym	**89.09**$_{\pm 1.57}$	58.88$_{\pm 3.30}$	67.72$_{\pm 2.92}$	57.43$_{\pm 0.23}$	66.95$_{\pm 0.12}$
QTL$_{\text{VT}}$	$\mathcal{P}_1$ + Synonym	88.47$_{\pm 0.80}$	69.43$_{\pm 3.89}$	78.42$_{\pm 2.75}$	61.30$_{\pm 4.36}$	71.92$_{\pm 3.16}$
	$\mathcal{P}_2$ + Synonym	**89.18**$_{\pm 0.63}$	70.58$_{\pm 1.49}$	**79.75**$_{\pm 0.81}$	65.30$_{\pm 3.67}$	75.32$_{\pm 3.15}$
	Sci-ZSEL + Synonym	87.99$_{\pm 1.47}$	**71.25**$_{\pm 2.58}$	79.49$_{\pm 1.65}$	**65.58**$_{\pm 2.62}$	**75.76**$_{\pm 2.03}$
QTL$_{\text{LPT}}$	$\mathcal{P}_1$ + Synonym	92.56$_{\pm 0.44}$	79.36$_{\pm 1.44}$	85.26$_{\pm 0.97}$	81.02$_{\pm 0.28}$	85.93$_{\pm 0.06}$
	$\mathcal{P}_2$ + Synonym	91.79$_{\pm 0.56}$	75.85$_{\pm 2.85}$	82.68$_{\pm 2.29}$	79.27$_{\pm 0.71}$	84.80$_{\pm 0.45}$
	Sci-ZSEL + Synonym	**92.66**$_{\pm 0.69}$	**80.70**$_{\pm 0.21}$	**86.12**$_{\pm 0.12}$	**81.21**$_{\pm 2.29}$	**86.21**$_{\pm 1.38}$

Table 13: Construction-strategy ablation of retriever (Recall@64, PC-POS) and reranker (Recall@1, MRR) results. We compare pseudo-pair construction strategies $\mathcal{P}_1$, $\mathcal{P}_2$, and their union Sci-ZSEL, each combined with curated synonyms. **Bold** indicates the best result per column within each dataset.

Ontology	Definition	Generated name
NCBI	Autosomal dominant HEREDITARY CANCER SYNDROME in which a mutation most often in either BRCA1 or BRCA2 is associated with a significantly increased risk for breast and ovarian cancers.	Hereditary Breast and Ovarian Cancer Syndrome
	The presence in a cell of two paired chromosomes from the same parent, with no chromosome of that pair from the other parent. This chromosome composition stems from non-disjunction (NONDISJUNCTION, GENETIC) events during MEIOSIS. The disomy may be composed of both homologous chromosomes from one parent (heterodisomy) or a duplicate of one chromosome (isodisomy).	Uniparental Disomy
	Acquired, familial, and congenital disorders of SKELETAL MUSCLE and SMOOTH MUSCLE.	Muscular Diseases
BC5CDR	A benzamide derivative that is used as a dopamine antagonist.	Tiapride Hydrochloride
	Pathologic processes that affect patients after a surgical procedure. They may or may not be related to the disease for which the surgery was done, and they may or may not be direct results of the surgery.	Postoperative Complications
	Used with drugs and chemicals for experimental human and animal studies of their ill effects. It includes studies to determine the margin of safety or the reactions accompanying administration at various dose levels. It is used also for exposure to environmental agents. Poisoning should be considered for life-threatening exposure to environmental agents.	toxicity
CMO	Any measurement that deals with the process or development of a disease state, for example, the onset, progression or severity of the disease or its symptoms.	Disease process measurement
	The maximum arterial pressure within the cardiac cycle, i.e. at the point at which the heart is in its maximal state of contraction. This is the time when the blood is forced from the ventricles of the heart into the pulmonary artery and the aorta.	Systolic blood pressure
	Total distance around the body at the region of the body lateral to and including the hip joint or coxa.	Hip circumference
VT	Any measurable or observable characteristic related to the physical magnitude of the portion of the body containing the brain and organs of sight, hearing, taste, and smell.	Head size trait
	The distance between the surfaces of the superficial layer of solid, hard bone that covers spongy bone.	Compact bone thickness
	The total proportion, quantity, or volume in milk of all inorganic elements or compounds that have importance in body functions.	Milk total mineral amount
LPT	Any measurable or observable characteristic related to the proportion or amount in milk of the alpha fraction of the casein phosphoprotein.	Milk alpha-casein content
	Any measurable or observable characteristic related to the relative heaviness of the skeletal tissue of an animal following slaughter and removal of the head, digestive tract, internal organs, and possibly the hide and/or feet.	Dressed carcass bone weight
	Any measurable or observable characteristic related to the relative heaviness of the muscle tissue surrounding the femur of a bird.	Thigh muscle weight

Table 14: Three in-context examples used in the entity generation prompt for each ontology. Definitions are taken verbatim from the respective source ontology.