BioELX: Context-Aware Cross-lingual Biomedical Entity Linking without Task-Specific Supervision

Anonymous ACL submission

Abstract

Cross-lingual biomedical entity linking (BEL) maps mentions in any language to unique identifiers in a biomedical knowledge base, supporting clinical and biomedical NLP applications. We identify two issues affecting current systems. First, the UMLS (Bodenreider, 2004) aliases used to train cross-lingual BEL retrievers are heavily skewed toward English, so retrievers generalize poorly to non-English mentions. Second, although context is often necessary for disambiguation, naively injecting context into retrievers trained only to align aliases severely degrades retrieval. We propose BioELX, a retrieve–rerank framework that addresses both issues. For retrieval, we continue training $\text{SapBERT}_{\text{multi}}$ (Liu et al., 2021b) using *Wikidata-derived cross-lingual alias supervision*, forming shared concept neighborhoods across languages. For reranking, we adapt pretrained LLM rerankers to entity linking through *mention-anchored prompting*, which marks the target mention so that rerankers score candidates with respect to the intended mention rather than other salient tokens in the context. Experiments show that BioELX achieves new state-of-the-art results on four cross-lingual BEL benchmarks, improving Recall@1 by 4.8 to 18.2 percentage points over prior best results, *without any task-specific BEL annotations*.

1 Introduction

Entity linking (EL) maps an ambiguous text span called a *mention* in a given text to the corresponding *entity* in a *knowledge base (KB)*, which stores entities as *unique IDs* associated with names, aliases, and optional metadata. Biomedical entity linking (BEL) is a domain-specific variant of EL, where mentions refer to biomedical entities such as diseases, drugs, symptoms, etc. Cross-lingual biomedical entity linking extends BEL to the multilingual setting, where mentions may appear in any language but must be linked to the same KB entities. In the example sentence "*Une réduction du nombre de globules rouges peut entraîner des symptômes tels que fatigue ou* ***essoufflement***."[1] in Fig. 1, the mention *essoufflement* should be linked to CUI C0013404, which has multilingual aliases such as *dyspnea* and *shortness of breath* in English, and *Sesak napas* in Indonesian. *CUI* stands for *concept unique identifier* and denotes IDs in the UMLS (Bodenreider, 2004), a comprehensive biomedical KB. However, KBs do not always contain an alias in the mention language: here, CUI C0013404 has no French alias, so the linker must connect the French mention to the correct concept through cross-lingual evidence.

Monolingual EL systems commonly adopt a two-stage design: *candidate retrieval* followed by *candidate reranking* (Wu et al., 2020; Cao et al., 2021; Orlando et al., 2024). This design balances efficiency and accuracy. The lightweight *retriever* efficiently selects a top-k candidate set from a KB over a precomputed entity embedding index, aiming for broad candidate coverage that includes the gold entity. A more expressive *reranker* then leverages richer information, such as mention context, to choose the final prediction. Reranking is more accurate but also more computationally intensive, because many rerankers jointly score each mention–candidate pair, requiring a separate forward pass per pair and limiting the use of pre-computed entity representations (Wu et al., 2020). Moreover, training such rerankers typically requires a substantial number of task-specific EL annotations.

In cross-lingual BEL, many existing systems are built on $\text{SapBERT}_{\text{multi}}$ (Liu et al., 2021b) as the retrieval component. Some extend or fine-tune it for specific languages and domains (Sakhovskiy et al., 2024; Mustafa et al., 2024; Fierens et al., 2025); others add a second-stage disambiguation model trained on task-specific BEL annotations (Zhu et al.,

[1]Translation: A reduction in the number of red blood cells can lead to symptoms such as fatigue or shortness of breath.

2023; Borchert et al., 2023; Donoso et al., 2024), which are scarce for many languages.

We identify two issues with current cross-lingual BEL models. First, the retriever $\text{SapBERT}_{\text{multi}}$ generalizes poorly to non-English mentions because its training resources provide limited multilingual evidence. This is apparent in the alias distribution used to train $\text{SapBERT}_{\text{multi}}$, where 69.6% of the aliases are written in English, while the remaining 30.4% are spread across 22 other languages (Liu et al., 2021b). This imbalance causes the learned representations to favor English surface forms, so non-English mentions are systematically drawn toward lexically similar but semantically incorrect entities. As shown in Fig. 3, the French mention *essou**ffle**ment* should be linked to the CUI C0013404 (*Dyspnea*), but $\text{SapBERT}_{\text{multi}}$ retrieves instead the lexically similar but semantically different concept *E**ffle**urage* ('light massage', CUI C0203907), likely because both strings share the rare substring **-ffle**.

This motivates our first hypothesis (H1): broader cross-lingual alias coverage should reduce the retriever's reliance on string overlap and enable non-English mentions to be linked through cross-lingual evidence. We test this hypothesis with *Wikidata-derived cross-lingual alias supervision*, which provides broader multilingual alias coverage for retriever training by mapping Wikidata aliases to biomedical concepts. §3.3 details the alias extraction, mapping, and training procedure.

A second issue is that contextual disambiguation is difficult to incorporate without sacrificing retrieval quality or relying on task-specific supervision. *Alias-based retrievers* such as $\text{SapBERT}_{\text{multi}}$ are trained to match aliases of the same concept, not to encode a mention with its context. Applying them to inputs that combine the mention with its context can distort the learned alias representation, shifting it away from the underlying concept and severely degrading retrieval. This effect is confirmed by our diagnostic experiment in Appendix D: adding context reduces $\text{SapBERT}_{\text{multi}}$ Recall@1 by more than 50 percentage points on both tested benchmarks. A standard approach to incorporate context is to use a dedicated context-aware reranker, but training such a reranker requires high-quality expert annotations that link mention spans to unambiguous entity identifiers. Creating such data is costly and such datasets are scarce, particularly for low-resource languages.

Pretrained LLM rerankers offer a training-free alternative because they are trained to compare query–document relevance, enabling context-aware candidate comparison without BEL-specific reranker training. We use rerankers rather than generative LLMs because entity generation can produce hallucinated or out-of-KB strings (Ji et al., 2023), whereas rerankers score candidates directly.

However, applying off-the-shelf LLM rerankers to EL introduces a task mismatch. These models are trained for general query–document relevance, not for entity disambiguation at a specific mention span. They may therefore focus on salient contextual entities instead of the target mention, which we call the *focus-mismatch problem*. For example, when reranking candidates for the mention *essouffflement* in Fig. 2, a reranker may attend to *globules rouges* and incorrectly prefer *Red Blood Cells* over the correct entity *Dyspnea*. This leads to our second hypothesis (H2): pretrained LLM rerankers can provide context-aware reranking without supervised BEL annotations, but must be adapted to score candidates with respect to the target mention rather than the whole context. We instantiate this adaptation with *mention-anchored prompting*, which marks the target mention in the input so that LLM rerankers score candidates against it. §3.4 details this adaptation.

Our contributions are: **1)** BioELX[2], a cross-lingual biomedical entity linking framework that can be applied to new benchmarks without training on any BEL-annotated data. BioELX achieves new state-of-the-art results on four cross-lingual benchmarks, improving the previous best by **+18.2** R@1 on XL-BEL (Liu et al., 2021b), **+5.5** on EMEA, **+4.8** on Patent (Kors et al., 2015), and **+8.4** on WikiMed-DE (Wang et al., 2023). Gains are especially pronounced for low-resource languages (Turkish **+18.2**, Korean **+21.1**, and Thai **+27.1** on XL-BEL); **2)** a Wikidata-derived cross-lingual resource for biomedical retrieval, providing broad multilingual concept-level supervision and **3)** a systematic study of how to adapt pretrained LLM rerankers to EL, that identifies mention-anchored prompting as a simple, training-free recipe for context-aware reranking. The architecture of BioELX is shown in Fig. 1. Following prior work (Wu et al., 2020; Agarwal et al., 2022; Sakhovskiy et al., 2024), we assume that mention spans are given and focus on linking each mention to a KB entity (see §3.1 for details).

[2]Code and resources will be released upon publication.

2 Related Work

2.1 Biomedical Entity Linking

BEL systems commonly follow a two-stage retrieve-then-rerank pipeline. Across both stages, prior work faces a recurring trade-off: methods that exploit mention context for disambiguation require task-specific BEL annotations, while annotation-free methods rely on KB aliases alone and ignore context.

Retrieval. Two paradigms dominate: the first trains retrievers on KB aliases with a contrastive objective, mapping different surface forms of the same concept to nearby embeddings (Sung et al., 2020; Liu et al., 2021a; Yuan et al., 2022b; Garda and Leser, 2024). It uses only KB aliases as supervision and does not condition on mention context. The second paradigm trains bi-encoders on annotated mention–entity pairs (Wu et al., 2020), encoding the mention in its context and the entity into a shared space (Agarwal et al., 2022; Xu et al., 2023; Lin et al., 2024); this exploits context but requires task-specific BEL annotations.

Reranking. Rerankers, in contrast, all incorporate mention context and differ mainly in how candidates are compared. Many score each mention–candidate pair independently, jointly encoding the mention context and one candidate entity to assign a relevance score (Agarwal et al., 2022; Sanz-Cruzado and Lever, 2025). Others place the mention context and several candidates in a shared input so they can be compared directly: Prompt-BioEL (Xu et al., 2023) and BioELQA (Lin et al., 2024) cast linking as a multiple-choice template, scored by predicting which candidate fills a masked slot in the template for Prompt-BioEL and using a generator to output the chosen candidate for BioELQA. In either way, training relies on task-specific BEL annotations.

Cross-lingual BEL. Liu et al. (2021b) introduced the XL-BEL benchmark and $\text{SapBERT}_{\text{multi}}$, a multilingual alias-based retriever trained contrastively on UMLS aliases, which has become the standard backbone for cross-lingual BEL. Many subsequent systems build directly on it. On the retrieval side, Mustafa et al. (2024) and Fierens et al. (2025) fine-tune $\text{SapBERT}_{\text{multi}}$ for German and French, and Sakhovskiy et al. (2024) add KB concept relations as a training signal. A complementary line of work adds a disambiguation stage after retrieval: Con2GEN (Zhu et al., 2023) trains a generative model to produce the target entity name, and ClinLinker (Donoso et al., 2024) trains a cross-encoder over mention context and candidate. Yet, retriever-side approaches still inherit the English-skewed alias supervision of UMLS and are typically tuned for a single language, while a disambiguation stage depends on BEL annotations that are scarce for most languages. BioELX targets both gaps: it broadens alias supervision with cross-lingual Wikidata aliases and adds context-aware reranking without BEL annotations (§3).

2.2 LLMs in Entity Linking

Prior work has used LLMs for EL in two main ways. The first treats the LLM as an auxiliary component of an existing EL pipeline, e.g., rewriting or augmenting mention context (Xin et al., 2025; Vollmers et al., 2025) or simplifying mentions to better match KB aliases (Borchert et al., 2024). The second reformulates EL as a generation task, where the LLM directly generates the target entity name or identifier, optionally constrained to a set of retrieved candidates (Xiao et al., 2023; Ding et al., 2024; Xie et al., 2024). However, given the diversity of entity aliases, generation-based linking is prone to instruction-following failures and hallucinated or out-of-KB outputs (Ji et al., 2023).

A third direction, largely unexplored for EL, is to use off-the-shelf LLM rerankers - models trained on large-scale query-document relevance data and widely used as rerankers in information retrieval (Chen et al., 2024; Wang et al., 2025; Zhang et al., 2025). Unlike generation models, rerankers score retrieved candidates and therefore cannot produce out-of-KB outputs; unlike pipeline-augmentation methods, they directly compare the mention with candidate entities. They are, however, trained for general query-document relevance rather than for mention-level disambiguation. BioELX adapts pre-trained LLM rerankers to cross-lingual BEL without task-specific BEL annotations (§3.4).

3 Methodology

3.1 Problem Formulation

Let $T = (t_1, \ldots, t_N)$ be a biomedical text of N tokens. We assume that a set of mentions $M = \{m_1, \ldots, m_{|M|}\}$, has already been identified in T, where each mention $m_i = (t_{m_{i,1}}, \ldots, t_{m_{i,n_i}})$ is a contiguous span of n_i tokens in T that refers to a biomedical concept. The target knowledge

base $\mathcal{K} = \{e_1, \ldots, e_{|\mathcal{K}|}\}$ contains $|\mathcal{K}|$ entities; each entity e has a unique identifier, an alias set $\mathcal{A}(e) = \{a_1, \ldots, a_{|\mathcal{A}(e)|}\}$ of surface forms in one or more languages, and optional metadata such as semantic types and textual descriptions.

Cross-lingual biomedical entity linking maps each mention to its corresponding KB entity. A mention may be written in any language, and that language need not be covered by the aliases of its gold entity in $\mathcal{K}$; this mismatch between the mention language and the KB is the defining difficulty of the cross-lingual setting. We assume each mention links to exactly one entity in $\mathcal{K}$. Formally, the task is to determine a linking function f such that $f(m, T, \mathcal{K}) = e$ for each mention $m \in M$ and its gold entity $e \in \mathcal{K}$.

3.2 BioELX Framework

Figure 1 shows the overall pipeline of BioELX, which comprises the BioELX Retriever for cross-lingual candidate retrieval and the BioELX Reranker for context-aware reranking.

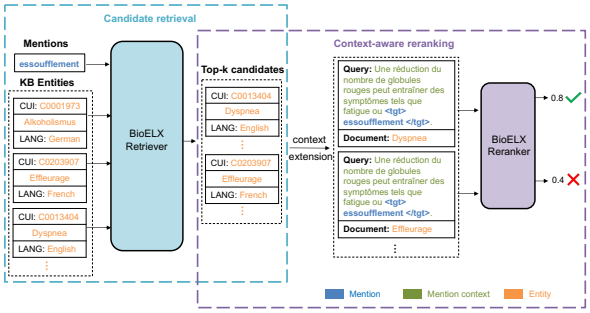

Figure 1: The two-stage BioELX pipeline, illustrated for a single mention. The retriever embeds the mention and multilingual entity aliases into a shared space, and returns the top-k aliases with associated entities by cosine similarity. For each candidate alias, the reranker scores a *mention-anchored query* — the mention and its surrounding context — against a document built from that alias, and selects the entity associated with the highest-scoring alias. Larger version in Figure 4.

Candidate retrieval. A single encoder $g(\cdot)$ embeds both mentions and aliases in a shared space. For efficiency, we embed every alias of every entity $e \in \mathcal{K}$ offline and index the resulting vectors; at inference time, we embed only the input mention m. Each alias is scored by its cosine similarity to the mention,

$$s_{\text{ret}}(m, a) = \cos(g(m), g(a)) \tag{1}$$

and we take the k highest-scoring aliases as the candidate set $C_k(m)$, each paired with its associated entity.

Context-aware reranking. Each candidate alias $a \in C_k(m)$ is then rescored in context. We construct a query–document input $\Phi\big(Q(m, T), D(a)\big)$, where the query $Q(m, T)$ is built from the mention and its surrounding context, and the document $D(a)$ contains an alias in $C_k(m)$ (see Figure 2). A pretrained LLM reranker R_θ assigns a relevance score

$$s_{\text{rank}}(m, a \mid T) = R_\theta\big(\Phi\big(Q(m, T), D(a)\big)\big) \in \mathbb{R}. \tag{2}$$

and BioELX predicts the highest-scoring alias,

$$a_{\text{final}} = \arg\max_{a \in C_k(m)} s_{\text{rank}}(m, a \mid T). \tag{3}$$

and outputs its associated entity. §3.4 specifies how BioELX constructs the mention-anchored query $Q(m, T)$ and candidate document $D(a)$ for LLM rerankers. Appendix E.3 describes the scoring functions used by the evaluated rerankers, including confidence-based scoring for QWEN3-Reranker (Zhang et al., 2025), cosine-similarity scoring for Jina-Reranker (Wang et al., 2025), and logit-based scoring for BGE-Reranker (Chen et al., 2024).

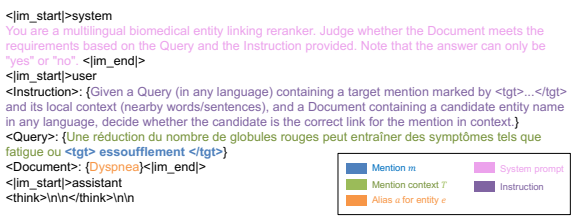

Figure 2: Prompt format for the BioELX Reranker, instantiated with the QWEN3-Reranker template (black text). The **query** contains the biomedical text T with the target mention m marked by `<tgt>`...`</tgt>`, and the **document** contains an alias a of the candidate entity e, so the reranker scores each mention-candidate pair. Our "document" is intentionally short: many KB entities lack rich textual metadata, so we provide only the entity alias. See larger version in Fig. 5.

3.3 Wikidata-derived Cross-lingual Retrieval

To address the issue of $\text{SapBERT}_{\text{multi}}$ discussed in Section 1, we construct *Wikidata-derived cross-lingual alias supervision*: instead of relying only on UMLS aliases, we use Wikidata-derived multilingual aliases to create cross-lingual concept-level

alias groups for retriever training. The intuition is that aliases of the same biomedical concept in different languages can serve as cross-lingual bridges in the embedding space. By pulling these aliases together during contrastive training, the retriever learns shared multilingual concept neighborhoods, so an unseen non-English mention can be drawn toward the correct concept rather than toward a lexically similar but semantically unrelated entity. We instantiate this idea by extracting multilingual aliases from Wikidata (Vrandecic and Krötzsch, 2014), mapping them to UMLS, and continuing contrastive training from SapBERT$_{\text{multi}}$. Figure 3 provides an illustrative example. For the French mention *essoufflement*, Wikidata-derived multilingual aliases provide broader cross-lingual evidence that helps place the correct entity closer to this mention in the embedding space, so it appears among the retrieved top-k candidates.

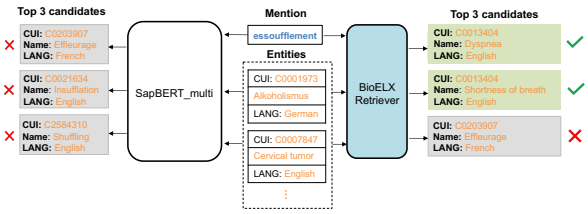

Figure 3: An example of retrieval. For the French mention *essoufflement*, SapBERT$_{\text{multi}}$ retrieves top candidates based on surface similarity but misses the correct concept. Our retriever, enhanced with Wikidata-derived multilingual aliases, retrieves candidates that include the correct CUI (Dyspnea) among the top results. See larger version in Fig. 6.

Constructing cross-lingual alias supervision from Wikidata. To provide multilingual alignment supervision beyond the English-skewed aliases in UMLS, we derive cross-lingual aliases from a recent Wikidata dump[3] and map them to UMLS CUIs through Wikidata identifiers (QIDs). Here, an *alias* denotes a language-specific surface form associated with a QID.

Concretely, we parse the Wikidata `item_per_site` dump to extract triples of the form (q, a, ℓ), where q is a QID, a is a language-specific alias, and ℓ denotes the corresponding language key. A single QID can have aliases in multiple languages. For example, `(3358860, 'Ovalocyte', frwiki)` indicates that the QID 3358860 has the French alias `'Ovalocyte'`. We collect all aliases in all languages for a given QID.

We then map each QID to a UMLS CUI by querying the Wikidata SPARQL endpoint[4] for items annotated with the property `P2892, UMLS CUI`[5]. Attaching this mapping to the extracted triples yields four-tuples (q, a, ℓ, c), where c is the associated CUI, e.g., `(3358860, 'Ovalocyte', frwiki, C0221280)`. We retain only tuples with a valid CUI mapping. To ensure that our retrieval gains reflect zero-shot generalization to unseen mentions rather than memorization of evaluation surface forms, we discard any tuple whose alias string exactly matches a mention string in our evaluation benchmarks. Finally, we group aliases by CUI to form cross-lingual positives for contrastive training.

We obtain $3,823,814$ aliases across 597 languages for $537,848$ unique UMLS CUIs. English accounts for only 6.3% of these aliases, while the nine non-English languages covered by XL-BEL together account for 11.3%. This sharply contrasts with the UMLS-derived aliases used to train SapBERT$_{\text{multi}}$, where English accounts for 69.6% (Liu et al., 2021b). Detailed statistics are provided in Appendix C.

Contrastive Training on Wikidata-Derived Aliases. Starting from SapBERT$_{\text{multi}}$[6], we continue training the encoder $g(\cdot)$ on Wikidata-derived cross-lingual aliases with the same contrastive learning objective, expanding multilingual alias coverage beyond UMLS and mitigating the English bias in alias supervision. Given a mini-batch $\mathcal{B} = \{(x_i, y_i)\}_{i=1}^{N}$, where x_i is an alias string and y_i is its CUI label, we define for the *anchor alias* x_i the in-batch positive and negative sets as

$$\mathcal{P}(x_i) = \{x_j \mid (x_j, y_j) \in \mathcal{B},\ x_j \neq x_i,\ y_j = y_i\} \tag{4}$$

$$\mathcal{N}(x_i) = \{x_j \mid (x_j, y_j) \in \mathcal{B},\ x_j \neq x_i,\ y_j \neq y_i\} \tag{5}$$

Each anchor x_i has at least one same-CUI positive in the batch (typically an alias in a different language); all remaining items in $\mathcal{B}$ serve as candidate negatives. We then form training triplets (x_i, x_{i^+}, x_{i^-}), where $x_{i^+} \in \mathcal{P}(x_i)$ and $x_{i^-} \in \mathcal{N}(x_i)$. Following SapBERT$_{\text{multi}}$ (Liu

[3] https://dumps.wikimedia.org/, version 2025-12-01.

[4] https://query.wikidata.org/

[5] https://www.wikidata.org/wiki/Property:P2892

[6] SapBERT-UMLS-2020AB-all-lang-from-XLMR-large

et al., 2021a), we apply in-batch hard mining and retain only *hard triplets* - those in which the negative alias is close to the anchor relative to the positive, namely those satisfying

$$\|g(x_i)-g(x_{i^+})\|_2+\lambda \;\geq\; \|g(x_i)-g(x_{i^-})\|_2, \quad (6)$$

for a predefined margin λ. Let $\mathcal{P}_i$ and $\mathcal{N}_i$ denote the indices of the positives and negatives that survive hard mining for x_i, and let $S_{ij} = \cos(g(x_i), g(x_j))$. We optimize the multi-similarity (MS) loss (Wang et al., 2019):

$$\mathcal{L}_{\text{ret}} = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{1}{\alpha}\log\left(1+\sum_{n\in\mathcal{N}_i}\exp\big(\alpha(S_{in}-\epsilon)\big)\right) + \frac{1}{\beta}\log\left(1+\sum_{p\in\mathcal{P}_i}\exp\big(-\beta(S_{ip}-\epsilon)\big)\right)\right] \quad (7)$$

where the hyperparameters α and β are scaling factors and ϵ is a similarity offset. This objective pulls together aliases with the same CUI while pulling apart aliases with different CUIs, encouraging a clustering of the vectors based on semantic instead of syntactic similarity, and improving language-agnostic retrieval.

3.4 Mention-Anchored Adaptation of LLM Rerankers

Many biomedical mentions are ambiguous and require context for correct disambiguation: in Figure 2 the mention *essoufflement* refers to the entity *Dyspnea*, not to the massage technique *Effleurage*, and only the surrounding text reveals that it is a symptom. Given the top-k candidate set returned by the retriever (§3.2), our goal is context-aware reranking without task-specific BEL annotations.

Why pretrained LLM rerankers. Pretrained LLM rerankers (Zhang et al., 2025; Wang et al., 2025; Chen et al., 2024) score retrieved candidates directly, without generating entity strings, thus avoiding the hallucinated or out-of-KB outputs of generation-based entity linking (Ji et al., 2023). We therefore formulate reranking by treating the mention in context as the query and each candidate's alias as the document.

The focus-mismatch problem. Off-the-shelf LLM rerankers are trained for general query–document relevance ranking, not mention-level disambiguation, and exhibit a *focus-mismatch problem*: the model may score against the most salient entity in the surrounding context instead of the target mention. Returning to the example in Fig. 2, *globules rouges* is medically salient in the context, but it is not the target of the linking decision. Appendix H.2 confirms this empirically - the original query–document formulation yields much weaker reranking performance than the mention-anchored formulation, indicating that off-the-shelf rerankers are not automatically aligned with the entity linking objective.

Mention-anchored prompting. To address the focus-mismatch, we explicitly mark the target mention in the query, signaling to the reranker which token span in the surrounding context should drive the scoring decision. The reranker input $\Phi\big(Q(m,T), D(a)\big)$ has two fields:

- The **query** $Q(m,T)$ is the biomedical text T with the target mention m wrapped in delimiters `<tgt>...</tgt>`.
- The **document** $D(a)$ is an alias of the candidate entity e.

Both fields, together with a task instruction defined by the prompt template, form the reranker input, as shown in Figure 2.

4 Experiments

We evaluate BioELX on four cross-lingual benchmarks - XL-BEL (Liu et al., 2021b), EMEA (Kors et al., 2015), Patent (Kors et al., 2015), and WikiMed-DE (Wang et al., 2023) - covering multiple languages, domains, and KB settings. Following prior work, we report Recall@1 and compare against strong multilingual and biomedical EL baselines. At inference time, the retriever returns the top-64 candidates, which are then reranked by QWEN3-Reranker (Zhang et al., 2025). Full dataset, baseline, and implementation details are provided in Appendix B. LLM reranker selection rationale is provided in Appendix E.

We additionally evaluate on MedMentions (Mohan and Li, 2019) as a large-scale English BEL transfer check, to examine whether Wikidata-derived alias supervision and mention-anchored reranking also benefit English BEL without supervised training. Since MedMentions is not a cross-lingual benchmark and most strong baselines are trained directly on its annotations, we report and discuss these results separately in Appendix F.2.

Dataset→	XL-BEL										Avg
Method↓	EN	ES	DE	FI	RU	TR	KO	ZH	JA	TH	
mBERT (Devlin et al., 2019)	29.1	11.9	5.2	1.6	5.1	8.0	0.6	3.6	9.1	0.9	7.5
mBART (Liu et al., 2020)	34.8	23.5	15.0	11.7	20.4	24.9	9.0	8.3	13.1	11.2	17.2
$\text{SapBERT}_{\text{multi}}$ (Liu et al., 2021b)	86.2	48.4	28.5	19.9	38.4	35.5	14.8	14.5	23.1	14.0	32.3
CODER (Yuan et al., 2022a)	85.9	51.4	31.0	20.3	37.3	35.4	1.4	14.4	22.6	2.4	30.2
mGENRE (Cao et al., 2022)	85.2	48.9	30.9	23.8	39.9	36.9	15.7	16.7	23.9	16.0	33.8
Con2GEN (Zhu et al., 2023)	86.6	52.7	30.8	24.4	39.8	40.5	17.4	17.1	25.6	19.2	35.4
BERGAMOT (Sakhovskiy et al., 2024)	78.6	58.2	33.2	22.9	37.3	41.9	18.5	18.9	25.4	21.5	35.6
$\text{SapBERT}_{\text{de}}$ (Mustafa et al., 2024)	83.0*	57.7*	40.1*	17.0*	33.5*	42.1*	17.4*	18.8*	23.8*	20.3*	35.4*
BioELX retriever (**ours**)	89.9	63.7	45.4	31.0	49.5	54.5	35.4	35.0	32.6	41.2	47.8
BioELX retriever+reranker (**ours**)	**91.2**	**69.6**	**53.0**	**35.9**	**56.6**	**60.3**	**39.6**	**44.5**	**39.1**	**48.6**	**53.8**

Table 1: R@1 (%) on XL-BEL. English (EN), Spanish (ES), German (DE), Finnish (FI), Russian (RU), Turkish (TR), Korean (KO), Chinese (ZH), Japanese (JA), Thai (TH). * indicates results reproduced by us. Other numbers are as reported by Zhu et al. and Sakhovskiy et al. The best results are bold, second best are underlined. For BioELX, we report the mean over three random seeds; standard deviations are provided in Appendix F.1.

Dataset→	EMEA				Avg	Patent		Avg	WikiMed-DE	
Method↓	ES	FR	NL	DE		FR	DE		DE	DE-wiki
mBERT (Devlin et al., 2019)	4.4	10.0	4.8	4.0	5.8	7.0	4.6	5.8	21.6*	70.1*
mBART (Liu et al., 2020)	35.6	30.9	29.0	29.9	31.4	27.2	29.0	28.1	28.3*	70.4*
$\text{SapBERT}_{\text{multi}}$ (Liu et al., 2021b)	58.3	53.4	55.5	53.6	55.2	63.5	61.2	62.4	54.6*	75.6
CODER (Yuan et al., 2022a)	56.9	53.6	56.9	53.4	55.2	64.9	65.8	65.4	23.5*	74.5*
mGENRE (Cao et al., 2022)	56.5	54.8	49.3	52.7	53.3	61.1	58.6	59.9	31.8*	73.6*
Con2GEN (Zhu et al., 2023)	63.9	57.1	51.4	55.5	57.0	64.0	66.1	65.1	NA	NA
BERGAMOT (Sakhovskiy et al., 2024)	60.2*	58.2*	54.4*	53.9*	56.7*	68.7*	68.4*	68.6*	51.0*	77.2*
$\text{SapBERT}_{\text{de}}$ (Mustafa et al., 2024)	56.5*	53.4*	52.5*	57.7*	55.0*	65.5*	67.2*	66.4*	51.3*	**80.0**
BioELX retriever (**ours**)	60.0	55.9	56.2	55.8	57.1	67.0	68.9	68.0	57.0	78.1
BioELX retriever+reranker (**ours**)	**65.4**	**61.9**	**59.9**	**62.7**	**62.5**	**75.4**	**71.4**	**73.4**	**63.0**	79.8

Table 2: R@1 (%) on EMEA, Patent and WikiMed-DE. We include two KBs for evaluation on German: (1) UMLS 2020AA (DE), and (2) UMLS-Wikidata (Mustafa et al., 2024) (DE-wiki). ⋇ indicates reproduced results, while NA means the code is not fully available. Other numbers are as reported by Zhu et al. The best results are bold, second best are underlined. We report the mean over three random seeds; standard deviations are provided in Appendix F.1.

4.1 Main Results

Tables 1 and 2 report R@1 on XL-BEL, EMEA, Patent, and WikiMed-DE. BioELX achieves the best overall performance across all four benchmarks, establishing a new state-of-the-art for cross-lingual EL.

Does cross-lingual alias supervision improve retrieval? Across the four benchmarks, the BioELX retriever substantially outperforms prior methods, confirming our first hypothesis that broader cross-lingual alias coverage helps retrieval for mentions in languages underrepresented in $\text{SapBERT}_{\text{multi}}$'s training data. The largest gains appear on XL-BEL: 47.8 average R@1 vs. 32.3 ($\text{SapBERT}_{\text{multi}}$) and 35.6 (BERGAMOT). Improvements are consistent across all XL-BEL languages and especially strong in low-resource settings (e.g., TR: 54.5 vs. 42.1, TH: 41.2 vs. 21.5 against the previous best results). The improvements also transfer beyond XL-BEL to other biomedical text genres: on WikiMed-DE with UMLS 2020AA, the retriever reaches 57.0 R@1, the best result in this setting. On EMEA and Patent, where surface-form ambiguity is higher, retrieval-only gains are smaller but still competitive. Appendix G.2 analyzes benchmark ambiguity and explains why alias-trained retrieval helps most on XL-BEL and WikiMed-DE while EMEA and Patent require context-aware reranking.

Do the gains come from cross-lingual alias coverage? Yes: among the 1,416 mentions improved over $\text{SapBERT}_{\text{multi}}$, 82.3% have gold CUIs with aliases in other languages, while none have only same-language Wikidata coverage (Appendix G). Broader multilingual alias coverage yields larger gains, but even concepts without any Wikidata-derived aliases in the training set still improve,

suggesting that continued contrastive training also reshapes the global geometry of the embedding space, reducing the model's reliance on superficial lexical similarity. The retrieval gains are therefore driven by cross-lingual alias coverage rather than same-language alias memorization or sheer exposure to more alias strings, further supporting our first hypothesis.

How much does our mention-anchored LLM reranking help beyond retrieval? Substantially: our pretrained LLM reranker with mention-anchored prompting adds 5-6 R@1 points across benchmarks, supporting our second hypothesis that pretrained LLM rerankers, once adapted to focus on the target mention, can provide context-aware reranking without supervised BEL annotations. On XL-BEL it increases the average R@1 from 47.8 to 53.8, improving performance in every language. EMEA improves from 57.1 to 62.5 and Patent from 68.0 to 73.4, exceeding all prior results. On WikiMed-DE (UMLS 2020AA), reranking adds 6.0 R@1 points over the retriever. Once the correct entity is included in the candidate set, context-aware candidate comparison improves linking - especially where alias similarity alone is insufficient.

When does reranking help less? Reranking helps less when there is little ambiguity to begin with. Table 2 shows a special case: WikiMed-DE with the UMLS-Wikidata KB (Mustafa et al., 2024). This KB is specifically constructed from Wikidata for WikiMed-DE, so many benchmark mentions closely match the corresponding entity aliases. Surface-form retrieval is therefore already highly effective, and the gold entity is often ranked at the top by the retriever alone. Adding the reranker yields only a marginal gain (78.1 $\rightarrow$ 79.8, +1.7). Most real-world BEL settings rarely come with such a benchmark-specific KB. They usually link mentions to comprehensive KBs like UMLS, which provide broader entity coverage and are closer to clinical practice. These settings leave more unresolved ambiguity after retrieval, which explains the larger reranking gains on XL-BEL, EMEA, Patent, and WikiMed-DE with UMLS 2020AA.

4.2 Ablation Study

To understand how retrieval quality affects reranking, we vary the retrievers (BM25, mBART, BERGAMOT, and the BioELX retriever) while keeping the reranker fixed. The results show that retrieval recall, R@k, caps the reranked R@1, so stronger retrieval leads to better overall performance. The BioELX retriever performs best among the evaluated retrievers, as reported in Appendix I.1.

We also examine whether adding entity metadata - such as entity type or description - improves reranking. As shown in Appendix I.2, this additional metadata does not improve performance.

Finally, three ablations in Appendix H provide additional evidence for H2. Appendix H.1 shows that mention-anchored prompting consistently outperforms lightweight supervised fine-tuning using LoRA adapters (Hu et al., 2022). Appendix H.2 compares the original query–document formulation with mention-anchored prompting across LLM reranker families and model sizes, showing that explicitly marking the target mention consistently mitigates the focus-mismatch problem across different LLM rerankers. Appendix H.3 analyzes mention-marker design, showing that the `<tgt>` marker performs best among those tested.

5 Conclusion

BioELX improves cross-lingual BEL without task-specific annotations by addressing two key challenges: limited multilingual evidence for non-English retrieval, and context-aware disambiguation without supervised BEL training. The gains stem from two complementary contributions. First, we show that Wikidata-derived cross-lingual alias supervision creates shared multilingual concept neighborhoods in the embedding space, enabling zero-shot generalization to unseen mentions through cross-lingual alias bridges; this mechanism alone yields +15.5 average R@1 points on XL-BEL over $\text{SapBERT}_{\text{multi}}$. Second, BioELX is, to our knowledge, among the first frameworks to adapt pretrained LLM rerankers to entity linking without task-specific BEL supervision. We identify a focus-mismatch problem when off-the-shelf LLM rerankers are applied to EL and propose mention-anchored prompting as a training-free solution that consistently improves results beyond retrieval alone across all benchmarks.

Limitations

While BioELX substantially improves cross-lingual BEL without task-specific annotations, the approach has the following limitations.

First, the reranker is constrained by the entity-side information available at inference time. In our main setting, the candidate document mainly contains the entity alias, because many KB entities lack rich textual descriptions. When multiple CUIs share near-identical metadata, such as aliases and entity types, even a strong context-aware reranker may not have sufficient candidate-side evidence to distinguish them. Enriching candidate representations with reliable descriptions or graph-based relations may improve fine-grained disambiguation, but such metadata are sparse and unevenly available across languages and KBs.

Second, following prior work (Wu et al., 2020; Agarwal et al., 2022; Sakhovskiy et al., 2024), we assume mentions are already identified. Our framework, therefore, addresses entity linking rather than end-to-end entity recognition and linking. In practical multilingual biomedical systems, mention detection can introduce additional errors, especially for low-resource languages and morphologically complex mentions. Extending BioELX with robust multilingual mention detection is an important direction for future work.

Ethical Considerations and Licenses

BioELX is intended to support biomedical entity linking research and downstream biomedical NLP applications. As with other biomedical NLP systems, incorrect entity linking predictions may affect downstream analyses if used without validation. Real-world applications in clinical decision-making settings should therefore include domain-expert review.

BioELX uses Wikidata-derived aliases as cross-lingual supervision and QWEN3-Reranker-8B as the main LLM reranker. Wikidata is released under the Creative Commons CC0 License. The QWEN3 embedding and reranking model series, including QWEN3-Reranker-8B, is released under the Apache 2.0 License. These uses are consistent with the stated licenses of the corresponding artifacts. Our released artifacts will include license and usage information.

Other pretrained models, datasets, and knowledge bases used in our evaluation are cited in the paper and used in accordance with their stated terms of use.

References

Dhruv Agarwal, Rico Angell, Nicholas Monath, and Andrew McCallum. 2022. Entity Linking via Explicit Mention-Mention Coreference Modeling. In *Proc. NAACL-HLT 2022*, pages 4644–4658. ACL.

Olivier Bodenreider. 2004. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res.*, 32(Database-Issue):267–270.

Florian Borchert, Ignacio Llorca, Roland Roller, Bert Arnrich, and Matthieu-P. Schapranow. 2023. xmen: A modular toolkit for cross-lingual medical entity normalization. *JAMIA Open*, 8.

Florian Borchert, Ignacio Llorca, and Matthieu-P. Schapranow. 2024. Improving biomedical entity linking for complex entity mentions with LLM-based text simplification. *Database J. Biol. Databases Curation*, 2024.

Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. 2021. Autoregressive Entity Retrieval. In *ICLR 2021*. OpenReview.

Nicola De Cao, Ledell Wu, Kashyap Popat, Mikel Artetxe, Naman Goyal, Mikhail Plekhanov, Luke Zettlemoyer, Nicola Cancedda, Sebastian Riedel, and Fabio Petroni. 2022. Multilingual Autoregressive Entity Linking. *TACL*, 10:274–290.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 2318–2335. ACL.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proc. NAACL-HLT 2019*, pages 4171–4186, Minneapolis, MN, USA. ACL.

Yifan Ding, Qingkai Zeng, and Tim Weninger. 2024. ChatEL: Entity Linking with Chatbots. In *Proc. LREC-COLING 2024*, pages 3086–3097. ELRA/ICCL.

Fernando Gallego Donoso, Guillermo López-García, Luis Gascó Sánchez, Martin Krallinger, and Francisco J. Veredas. 2024. Clinlinker: Medical entity linking of clinical concept mentions in spanish. In *Proc. ICCS 2024*, volume 14836 of *LNCS*, pages 266–280. Springer.

Amaury Fierens, Alexandre Englebert, and Sébastien Jodogne. 2025. Translating UMLS Concepts to Improve Medical Entity Linking in French: A SapBERT-Based Approach. In *Intelligent Health Systems –*

From Technology to Data and Knowledge, volume 327 of *Studies in Health Technology and Informatics*, pages 870–871. IOS Press. PMID: 40380592.

Samuele Garda and Ulf Leser. 2024. BELHD: improving biomedical entity linking with homonym disambiguation. *Bioinformatics*, 40(8):btae474.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *ICLR 2022*. OpenReview.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv.*, 55(12):248:1–248:38.

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 ACL 2025*, pages 10714–10730. ACL.

Jan A. Kors, Simon Clematide, Saber A. Akhondi, Erik M. van Mulligen, and Dietrich Rebholz-Schuhmann. 2015. A multilingual gold-standard corpus for biomedical concept recognition: the Mantra GSC. *J. Am. Med. Inform. Assoc.*, 22(5):948–956.

Zhenxi Lin, Ziheng Zhang, Xian Wu, and Yefeng Zheng. 2024. Biomedical Entity Linking as Multiple Choice Question Answering. In *Proc. LREC-COLING 2024*, pages 2390–2396, Torino, Italia. ELRA/ICCL.

Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. 2021a. Self-Alignment Pretraining for Biomedical Entity Representations. In *Proc. NAACL-HLT 2021*, pages 4228–4238. ACL.

Fangyu Liu, Ivan Vulic, Anna Korhonen, and Nigel Collier. 2021b. Learning Domain-Specialised Representations for Cross-Lingual Biomedical Entity Linking. In *Proc. ACL-IJCNLP 2021 (Short)*, pages 565–574. ACL.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual Denoising Pre-training for Neural Machine Translation. *TACL*, 8:726–742.

Lajanugen Logeswaran, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, Jacob Devlin, and Honglak Lee. 2019. Zero-shot entity linking by reading entity descriptions. In *Proc. ACL 2019*, pages 3449–3460, Florence, Italy. ACL.

Sunil Mohan and Donghui Li. 2019. MedMentions: A Large Biomedical Corpus Annotated with UMLS Concepts. In *AKBC 2019*.

Faizan E. Mustafa, Corina Dima, Juan G. Díaz Ochoa, and Steffen Staab. 2024. Leveraging Wikidata for Biomedical Entity Linking in a Low-Resource Setting: A Case Study for German. In *Proc. ClinicalNLP Workshop 2024*, pages 202–207, Mexico City, Mexico. ACL.

Riccardo Orlando, Pere-Lluís Huguet Cabot, Edoardo Barba, and Roberto Navigli. 2024. ReLiK: Retrieve and LinK, Fast and Accurate Entity Linking and Relation Extraction on an Academic Budget. In *Findings of ACL 2024*, pages 14114–14132. ACL.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.

Andrey Sakhovskiy, Natalia Semenova, Artur Kadurin, and Elena Tutubalina. 2024. Biomedical Entity Representation with Graph-Augmented Multi-Objective Transformer. In *Findings of NAACL 2024*, pages 4626–4643, Mexico City, Mexico. ACL.

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, Viena, Austria. Association for Computational Linguistics.

Mujeen Sung, Hwisang Jeon, Jinhyuk Lee, and Jaewoo Kang. 2020. Biomedical Entity Representations with Synonym Marginalization. In *Proc. ACL 2020*, pages 3641–3650. ACL.

Daniel Vollmers, Hamada M. Zahera, Diego Moussallem, and Axel-Cyrille Ngonga Ngomo. 2025. Contextual Augmentation for Entity Linking using Large Language Models. In *Proc. COLING 2025*, pages 8535–8545, Abu Dhabi, UAE. ACL.

Denny Vrandecic and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85.

Feng Wang, Yuqing Li, and Han Xiao. 2025. jina-reranker-v3: Last but not late interaction for document reranking. *arXiv*, abs/2509.25085.

Xun Wang, Xintong Han, Weilin Huang, Dengke Dong, and Matthew R. Scott. 2019. Multi-similarity loss with general pair weighting for deep metric learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 5022–5030. Computer Vision Foundation / IEEE.

Yi Wang, Corina Dima, and Steffen Staab. 2023. WikiMed-DE: Constructing a Silver-Standard Dataset for German Biomedical Entity Linking using Wikipedia and Wikidata. In *Wikidata Workshop @ ISWC 2023*, volume 3640 of *CEUR Workshop Proc.* CEUR-WS.org.

Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. 2020. Scalable Zero-shot Entity Linking with Dense Entity Retrieval. In *Proc. EMNLP 2020*, pages 6397–6407. ACL.

Zilin Xiao, Ming Gong, Jie Wu, Xingyao Zhang, Linjun Shou, and Daxin Jiang. 2023. Instructed Language Models with Retrievers Are Powerful Entity Linkers. In *Proc. EMNLP 2023*, pages 2267–2282. ACL.

Yuzhang Xie, Jiaying Lu, Joyce C. Ho, Fadi B. Nahab, Xiao Hu, and Carl Yang. 2024. PromptLink: Leveraging Large Language Models for Cross-Source Biomedical Concept Linking. In *Proc. SIGIR 2024*, pages 2589–2593. ACM.

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 *Proc. CIKM 2025*, pages 3550–3559. ACM.

Zhenran Xu, Yulin Chen, and Baotian Hu. 2023. Improving Biomedical Entity Linking with Cross-Entity Interaction. In *AAAI 2023*, pages 13869–13877. AAAI.

Hongyi Yuan, Zheng Yuan, and Sheng Yu. 2022a. Generative Biomedical Entity Linking via Knowledge Base-Guided Pre-training and Synonyms-Aware Fine-tuning. In *Proc. NAACL-HLT 2022*, pages 4038–4048. ACL.

Zheng Yuan, Zhengyun Zhao, Haixia Sun, Jiao Li, Fei Wang, and Sheng Yu. 2022b. CODER: Knowledge-infused cross-lingual medical term embedding for term normalization. *J. Biomed. Informatics*, 126:103983.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 Embedding: Advancing Text Embedding and Reranking Through Foundation Models. *arXiv*, abs/2506.05176.

Liangyu Zhong, Fabio Rosenthal, Joachim Sicking, Fabian Hüger, Thorsten Bagdonat, Hanno Gottschalk, and Leo Schwinn. 2026. FOCUS: Internal MLLM representations for efficient fine-grained visual question answering. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Tiantian Zhu, Yang Qin, Qingcai Chen, Xin Mu, Changlong Yu, and Yang Xiang. 2023. Controllable Contrastive Generation for Multilingual Biomedical Entity Linking. In *Proc. EMNLP 2023*, pages 5742–5753. ACL.

Tiantian Zhu, Yang Qin, Ming Feng, Qingcai Chen, Baotian Hu, and Yang Xiang. 2024. BioPRO: Context-Infused Prompt Learning for Biomedical Entity Linking. *IEEE ACM Trans. Audio Speech Lang. Process.*, 32:374–385.

A Notation Summary

Symbol	Meaning
$T = (t_1, \ldots, t_N)$	Input biomedical text.
t_i	The i-th token in T.
M	Set of detected mentions.
m	mention
m_i	The i-th mention.
n_i	Length of mention m_i.
$\mathcal{K}$	Target biomedical knowledge base.
e	Entity in the KB.
e_i	the i-th Entity in the KB.
$f(m, T, \mathcal{K})$	Entity-linking function.
$\mathcal{A}(e)$	Alias set associated with entity e.
a	Alias.
$g(\cdot)$	Retriever encoder.
$\cos(\cdot)$	Cosine similarity.
$s_{\text{ret}}(m, a)$	Retrieval score between mention m and alias a.
$C_k(m)$	Top-k candidate alias set for the mention m.
k	Number of retrieved candidates.
$\Phi(Q(m, T), D(a))$	Reranker input.
$Q(m, T)$	Mention-anchored query.
$D(a)$	Candidate document.
R_θ	Pretrained LLM reranker.
$s_{\text{rank}}(m, a \mid T)$	Reranking score.
a_{final}	The final alias after reranking.
q	Wikidata item identifier.
ℓ	Language key.
c	UMLS concept unique identifier.
(q, a, ℓ)	Wikidata alias triple.
(q, a, ℓ, c)	Alias tuple mapped to UMLS.
$\mathcal{B} = \{(x_i, y_i)\}_{i=1}^{N}$	Mini-batch for contrastive training.
x_i	Alias string.
y_i	CUI label of x_i.
x_{i+}	positive alias for x_i.
x_{i-}	negative alias for x_i .
$\mathcal{P}(x_i)$	Positive alias set for x_i.
$\mathcal{N}(x_i)$	Negative alias set for x_i.
S_{ij}	Cosine similarity of x_i and x_j.
S_{ip}	Cosine similarity of x_i and positive alias x_p.
S_{in}	Cosine similarity of x_i and negative alias x_n.
λ	Hard-mining margin.
α, β	Scaling factors in MS loss.
ϵ	Similarity offset in MS loss.
$\mathcal{L}_{\text{ret}}$	Retrieval training loss.

Table 3: Notation used in the paper.

B Experimental Setup and Implementation

B.1 Datasets

We evaluate BioELX on five biomedical entity linking benchmarks, covering multiple languages, diverse domains, and different settings (i.e., from single sentences to paragraphs). Table 4 summarizes the number of mentions and unique entities per language for each benchmark.

- **XL-BEL** (Liu et al., 2021b): a multilingual benchmark covering 10 typologically diverse languages: English (EN), Spanish (ES), German (DE), Finnish (FI), Russian (RU), Turkish (TR), Korean (KO), Chinese (ZH), Japanese (JA), Thai (TH). Each language contains 1,000 annotated sentences extracted from Wikipedia, each with a single annotated mention.

	XL-BEL										EMEA				Patent		WikiMed-DE	MedMentions
	EN	ES	DE	FI	RU	TR	KO	ZH	JA	TH	ES	FR	NL	DE	FR	DE	DE	EN
Mentions	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	432	431	434	425	342	348	40,703	40,116
Entities	807	889	908	816	816	694	829	875	822	704	295	304	304	299	223	223	9,123	8,457

Table 4: Dataset statistics for XL-BEL, EMEA, Patent, WikiMed-DE and MedMentions across languages. For WikiMed-DE and MedMentions, we only report the statistics for their test set because we do not use them for training.

- **EMEA** (Kors et al., 2015): the text of this benchmark comes from the European Medicines Agency. It covers four languages (ES, FR, NL, DE) with sentence-level contexts and UMLS-linked mentions.

- **Patent** (Kors et al., 2015): text of this benchmark comes from the European Patent Office. It contains two languages, German and French, also annotated with UMLS-linked mentions.[7]

- **WikiMed-DE** (Wang et al., 2023): a large-scale, silver-standard German benchmark constructed from Wikipedia with UMLS-aligned entities; we report results on its test set containing 40,703 mentions.

- **MedMentions** (Mohan and Li, 2019) an English biomedical entity linking benchmark constructed from PubMed abstracts, with expert-linked mentions to UMLS CUIs. Since MedMentions is not cross-lingual and most strong baselines are trained on its annotations, we use it only as an additional English BEL transfer check. It covers a broad range of biomedical concepts and provides document-level contexts; we evaluate on the standard test split.

B.2 Knowledge Bases

To ensure comparability, we follow prior work (Liu et al., 2021b; Zhu et al., 2023; Mustafa et al., 2024) and use the same knowledge bases for evaluation:

- **UMLS 2020AA version** (Bodenreider, 2004); we use the full release, containing 3,376,510 entities. We use this KB to evaluate EMEA, Patent and WikiMed-DE (UMLS setting).

- **UMLS 2017AA version** (Bodenreider, 2004); we use the full release of this version, containing 2,595,474 entities to evaluate MedMentions, which was annotated using this UMLS version.

- **UMLS-399931** (Liu et al., 2021b): the KB provided with the XL-BEL benchmark. It is a subset of UMLS 2020AA, containing 62,094 entities and 399,931 aliases. We use this KB to evaluate XL-BEL.

- **UMLS–Wikidata** (Mustafa et al., 2024): a German KB constructed for WikiMed-DE by collecting German Wikidata items that are linked to a UMLS CUI. We use this KB only for evaluating WikiMed-DE under the UMLS–Wikidata setting.

B.3 Baselines

Following prior studies (Sakhovskiy et al., 2024; Zhu et al., 2023; Mustafa et al., 2024), we compare BioELX against representative baselines covering dense retrieval, alias-based biomedical retrieval, generative linking, KB-enhanced retrieval, and supervised biomedical EL. For the cross-lingual benchmarks, we compare against multilingual BEL systems. For MedMentions, we follow the evaluation setting in Kim et al. (2025) and compare against strong systems trained specifically on MedMentions.

Baselines for multilingual benchmarks.

- **mBERT** (Devlin et al., 2019): a multilingual Transformer encoder used as a dense retriever. Mentions and entity names are encoded into the same vector space, and linking is performed by nearest-neighbor search. Unlike biomedical alias-trained models, mBERT is not specialized for biomedical synonym alignment.

- **mBART** (Liu et al., 2020): a multilingual sequence-to-sequence model used as an encoder baseline for dense retrieval. We encode mentions and entity names with the same model and perform nearest-neighbor search in the embedding space.

[7] We follow the same data source as EMEA; the "Patent" subset is reported separately due to its distinct domain.

- **SapBERT$_{\text{multi}}$** (Liu et al., 2021b): a multilingual biomedical retriever trained with a contrastive objective on UMLS aliases. It learns to align aliases that share the same CUI and serves as the main alias-based retrieval baseline. BioELX builds on this representation but expands the cross-lingual alias supervision with Wikidata-derived aliases.

- **CODER** (Yuan et al., 2022a): a biomedical encoder that incorporates knowledge signals such as concept structure into contrastive representation learning. It is used as a dense retrieval baseline in the same nearest-neighbor linking setup.

- **mGENRE** (Cao et al., 2022): a multilingual generative entity linking model. Instead of retrieving candidates with dense similarity, it generates entity identifiers or names through constrained decoding.

- **Con2GEN** (Zhu et al., 2023): a cross-lingual biomedical entity linking framework that combines SapBERT$_{\text{multi}}$ retrieval with controllable generation based on mGENRE. It injects KB-side attributes (e.g., entity type) into the generation process to better disambiguate lexically ambiguous mentions. We do not report Con2GEN on WikiMed-DE because there is no fully reproducible code for this baseline.

- **BERGAMOT** (Sakhovskiy et al., 2024): a cross-lingual biomedical entity linking approach that incorporates structured UMLS knowledge (e.g., graph) for representation learning.

- **SapBERT$_{\text{de}}$** (Mustafa et al., 2024): A German-adapted model fine-tuned based on SapBERT$_{\text{multi}}$ for WikiMed-DE. We also report its results on other benchmarks to compare with our model, BioELX.

All baseline numbers in our main paper are reported by Zhu et al. and Sakhovskiy et al. except the numbers annotated with the symbol *. For the numbers annotated with *, we run the released checkpoints under the same evaluation protocol.

Baselines for MedMentions. We additionally evaluate on MedMentions as an English BEL transfer check. Except for SapBERT$_{\text{multi}}$, the following baselines are trained on the MedMentions training set, so the comparison is not supervision-matched with BioELX.

- **SapBERT$_{\text{multi}}$**: the same UMLS-alias-trained biomedical retriever used in the multilingual experiments. It provides a direct, retrieval-only comparison to assess whether BioELX improves over the original alias-based representation.

- **GenBioEL** (Yuan et al., 2022a): a supervised generative BEL model that casts linking as sequence generation, producing the entity name/identifier with constrained decoding over a KB-derived vocabulary space.

- **ArboEL** (Agarwal et al., 2022): a supervised retrieve–rerank framework trained on MedMentions, which trains a bi-encoder for candidate retrieval and a cross-encoder for reranking using mention context and entity text.

- **Prompt-BioEL** (Xu et al., 2023): a supervised retrieve–rerank method. It trains a bi-encoder retriever on mentions and entity names, and then applies a prompt-tuned reranker that jointly encodes the mention context with candidate entities to model cross-entity interaction.

- **GenBioEL+ANGEL** (Kim et al., 2025): a biomedical entity linking system fine-tuned based on GenBioEL with hard negative optimization.

All numbers for these approaches are reported in the original papers. These MedMentions baselines are optimized directly on the target benchmark, whereas BioELX does not use MedMentions supervision. We therefore use MedMentions only as an additional English transfer check rather than as the primary setting for our cross-lingual BEL claims.

B.4 Implementation Details

Continued Contrastive Training of BioELX retriever. We follow the SapBERT$_{\text{multi}}$ training setup and construct positive alias pairs within each CUI. To avoid over-representing CUIs with many aliases and to keep training efficient, we randomly sample at most 50 positive pairs per CUI. We use the margin $\lambda = 0.2$ for online hard example mining, and represent each input using the `[CLS]` embedding. Training runs for 5 epochs with a batch size of 256, a maximum sequence length of 25,

and a learning rate of 2e-5, and is optimized using AdamW with a weight decay of 0.01.

Inference. At inference time, we use BioELX Retriever to generate the top-64 candidate aliases with associated CUIs for each mention. We then apply the QWEN3-Reranker-8B model (Zhang et al., 2025) to score each (query, candidate-document) pair. The reranker operates in batches over candidates, producing a scalar relevance score for each candidate alias. Candidate aliases are re-sorted according to the reranker scores, and the final prediction is the CUI associated with the highest-scoring alias.

Metrics. Following prior work on EL (Logeswaran et al., 2019; Wu et al., 2020), we report Recall@k (R@k), i.e., the fraction of mentions for which the gold CUI appears in the top-k predictions. In our evaluation setting, each mention has a single gold CUI, so R@1 is numerically equivalent to Precision@1 and accuracy.

B.5 Reproducibility and Computational Cost

To support reproducibility, we will release the full codebase upon publication. This includes scripts for processing Wikidata dumps and mapping Wikidata aliases to UMLS CUIs via Wikidata properties, training the BioELX Retriever, and running inference with both the retriever and the LLM reranker. We will also release the processed Wikidata-derived aliases and upload the trained BioELX Retriever checkpoint to Hugging Face.

All main experiments are run on a machine with 8 NVIDIA A100 80GB GPUs. Training the retriever takes approximately 45 minutes per epoch. Processing Wikidata-derived aliases is CPU-based and does not require GPUs. At inference time, retrieval and reranking have different computational profiles. The retriever requires GPU computation to encode mentions and KB aliases, but KB alias embeddings can be precomputed once and reused across runs. This one-time preprocessing cost scales with the KB size: encoding the XL-BEL KB takes approximately 3 minutes, whereas encoding the full UMLS 2020 release with millions of entity names and aliases takes approximately 2 hours.

In our main setting, we retrieve and rerank the top-64 candidates per mention with QWEN3-Reranker-8B. On XL-BEL, which contains 10,000 mentions and therefore 640,000 mention–candidate pairs, reranking takes approximately 1 hour and 44 minutes. Since reranking cost scales with the number of mention–candidate pairs, reducing the candidate set can substantially reduce inference time.

C Statistics of Wikidata-Derived Alias Supervision

This section provides detailed statistics of our Wikidata-derived multilingual aliases used to train the BioELX retriever. Table 5 reports statistics at three stages of constructing the Wikidata-derived alias set. The *Raw Wikidata extraction* block shows the number of aliases, QIDs, and languages obtained directly from the Wikidata dump before any biomedical filtering. The *After CUI mapping* block reports the subset of Wikidata items that can be linked to UMLS through the `P2892` property, together with the resulting number of CUI-linked aliases, QIDs, CUIs, and languages. The *After removing benchmark aliases* block provides the final dataset after decontamination, with aliases that exactly match evaluation mention strings removed. We also report the average and median numbers of languages per CUI in the final training set, along with the end-to-end processing time.

Statistic	**Value**
Raw Wikidata extraction	
Wikidata dump version	2025-12-01
Dump file used	`wb_items_per_site.sql`
Dump file size	1.69 GB
Raw aliases extracted	96,593,429
Raw unique QIDs extracted	33,696,395
Raw languages covered	950
After CUI mapping	
QIDs with P2892 annotation	731,652
CUI-linked aliases after mapping	3,834,319
Unique QIDs after CUI mapping	534,156
Unique CUIs after mapping	537,848
Languages covered after mapping	597
After removing benchmark aliases	
Total aliases	3,823,814
Unique QIDs	534,131
Unique CUIs	537,823
Languages covered	597
Avg. languages per CUI	7
Median languages per CUI	7
English aliases coverage	6.34%
Processing time (end-to-end)	13.83 Min
Workstation	CPU only

Table 5: Overall statistics of the Wikidata-derived alias extraction, CUI mapping, decontamination, and final training set construction pipeline.

Dataset	Input	R@1	R@8	R@16	R@32	R@64
EMEA	Mention only	55.2	71.2	76.3	79.2	80.1
	Mention + context	5.3	14.9	18.8	22.5	22.5
Patent	Mention only	62.4	79.8	82.9	86.8	90.3
	Mention + context	3.2	10.8	14.4	18.3	24.8

Table 6: Effect of naively appending context to $\text{SapBERT}_{\text{multi}}$ retrieval on EMEA and Patent. Results are R@k (%).

D Diagnostic Experiments

As an initial diagnostic experiment, we test whether context can be directly incorporated into an alias-trained retriever. We use the off-the-shelf $\text{SapBERT}_{\text{multi}}$ checkpoint[8] and compare two retrieval inputs: the mention string alone, and the mention concatenated with its surrounding context. In both settings, the retrieval index is constructed from the same entity names and aliases, and only the input changes. Since we use a fixed pretrained checkpoint and do not train any model in this experiment, the retrieval results are deterministic and do not involve random seeds.

As shown in Table 6, naively appending context severely degrades retrieval at all cutoffs. On EMEA, R@1 drops from 55.2 to 5.3 and R@64 from 80.1 to 22.5. On Patent, R@1 drops from 62.4 to 3.2 and R@64 from 90.3 to 24.8. These results confirm that alias-trained retrievers are sensitive to non-alias context tokens. This supports our design choice of keeping retrieval at mention level and introducing context at the reranking stage.

E LLM Reranker Details and Analysis

E.1 Why LLM Rerankers

For context-aware reranking without task-specific BEL annotations, we adopt LLM rerankers rather than generation-based LLMs. Generation-based entity linking requires the model to produce an entity name or identifier, which can lead to hallucinated or out-of-KB outputs (Ji et al., 2023), especially given the diversity of biomedical aliases. In contrast, LLM rerankers score query–document pairs directly, using a relevance score, confidence score, or similarity score depending on the model. This allows us to rank candidate entities without generating entity strings.

[8] https://huggingface.co/cambridgeltl/SapBERT-UMLS-2020AB-all-lang-from-XLMR.

E.2 LLM Reranker Selection Criteria

We select rerankers according to three practical criteria. First, the reranker should be open-source, both to avoid additional usage costs and to support reproducible research. Second, it should provide a relevance score, confidence score, or similarity score for each query–document pair, rather than generating entity strings. This scoring-based formulation avoids generation-based failure modes such as hallucinated or out-of-KB outputs. Third, it should support multilingual ranking, which is essential for cross-lingual BEL. Representative rerankers satisfying these criteria include QWEN3-Ranker (Zhang et al., 2025), Jina-Reranker (Wang et al., 2025), and BGE-Reranker (Chen et al., 2024).

We use QWEN3-Reranker-8B as the main reranker in our primary experiments because it provides a strong open-source multilingual ranking interface and performs best among the evaluated rerankers in our transfer analysis. Appendix H.2 shows that mention anchoring improves all evaluated rerankers, while QWEN3-Reranker-8B obtains the strongest overall reranking results. We therefore use it for the main tables and use Jina-Reranker and BGE-Reranker as additional checks of whether mention anchoring transfers across reranker architectures.

E.3 Scoring Function Variants for Evaluated Rerankers

Mention-anchored prompting is independent of how the LLM reranker produces its scalar relevance score. Different open-source rerankers expose different scoring interfaces, summarized below. In all variants, the prompt format is identical: the query $Q(m, T)$ contains the biomedical text with the target mention wrapped in `<tgt>...</tgt>`, and the document $D(a)$ contains the candidate entity alias.

Confidence-based scoring function. QWEN3-Reranker (Zhang et al., 2025) is trained to output a binary "yes" or "no" decision indicating whether the document is relevant to the query. Following Zhang et al. (2025), the relevance score is the model's confidence in the "yes" token, obtained by extracting the logits over the vocabulary at the answer position, restricting to the "yes" and "no" tokens, applying softmax, and taking the probabil-

ity of "yes":

$$s^{\text{conf}}_{\text{rank}}(m, a \mid T) = \text{Softmax}\Big(l(\text{yes} \mid Q, D), l(\text{no} \mid Q, D)\Big)_{\text{yes}}. \quad (8)$$

where $l(\text{yes} \mid Q, D)$ and $l(\text{no} \mid Q, D)$ are logits for the next token being "yes" or "no" conditioned on the concatenated query–document input, respectively. This is the scoring form used in our main experiments.

Similarity-based scoring function. Jina-Reranker-v3 (Wang et al., 2025) places the query and all candidate documents in a single context window and extracts contextual embeddings from designated special-token positions. A two-layer MLP projects these embeddings into a shared ranking space, and the relevance score is the cosine similarity between the projected query embedding $\mathbf{q}$ for Q and the projected document embedding $\mathbf{d}$ for D:

$$s^{\text{sim}}_{\text{rank}}(m, a \mid T) = \cos\big(\mathbf{q}, \mathbf{d}\big) \in [-1, 1]. \quad (9)$$

Logit-based scoring function. BGE-Reranker-v2-Gemma (Chen et al., 2024) is a cross-encoder that jointly encodes the query–document pair and produces a single scalar relevance logit:

$$s^{\text{logit}}_{\text{rank}}(m, a \mid T) = \ell_{\text{yes}}\big(Q(m, T), D(a)\big) \in \mathbb{R}, \quad (10)$$

where ℓ_{yes} is the model's logit for the "yes" token. Higher logits indicate higher relevance. We use the raw logits without sigmoid normalization, since within-ranker ordering is sufficient for sorting candidates.

F Additional Result Details

F.1 Standard Deviations over Random Seeds

For BioELX, the main tables in Section 4.1 report mean R@1 over three random seeds. Table 7 provides the corresponding mean and standard deviation values. For BioELX Retriever, we train the retriever with three different random seeds and evaluate R@1 for each dataset/language in each run. We then compute the mean and standard deviation of the three R@1 values for each reported column. For BioELX, the reranker is fixed and not trained by us. We apply the same fixed reranker to the candidate sets produced by the three retriever runs, evaluate the resulting end-to-end R@1 for each run, and compute the mean and standard deviation across the three runs. All values are reported as mean (standard deviation) in R@1 percentage points.

Dataset	Lang.	BioELX Retriever	BioELX
XL-BEL	EN	89.9 (0.42)	91.2 (0.23)
XL-BEL	ES	63.7 (0.36)	69.6 (0.25)
XL-BEL	DE	45.4 (0.40)	53.0 (1.25)
XL-BEL	FI	31.0 (0.38)	35.9 (0.61)
XL-BEL	RU	49.5 (0.57)	56.6 (0.46)
XL-BEL	TR	54.5 (0.57)	60.3 (0.85)
XL-BEL	KO	35.4 (0.37)	39.6 (0.35)
XL-BEL	ZH	35.0 (0.49)	44.5 (0.38)
XL-BEL	JA	32.6 (0.40)	39.1 (0.46)
XL-BEL	TH	41.2 (0.85)	48.6 (1.10)
EMEA	ES	60.0 (0.48)	65.4 (0.35)
EMEA	FR	55.9 (0.61)	61.9 (0.46)
EMEA	NL	56.2 (1.51)	59.9 (0.39)
EMEA	DE	55.8 (0.27)	62.7 (0.76)
Patent	FR	67.0 (1.82)	75.4 (0.01)
Patent	DE	68.9 (0.33)	71.4 (1.29)
WikiMed-DE	DE	57.0 (0.29)	63.0 (0.65)
WikiMed-DE	DE-wiki	78.1 (0.12)	79.8 (0.30)
MedMentions	EN	53.3 (0.03)	60.1 (0.47)

Table 7: R@1 (%) over three random seeds. Each value is reported as mean (standard deviation). **Lang.** denotes language. For **BioELX Retriever**, the standard deviation is computed over three independently trained retrievers. For **BioELX**, the reranker is fixed; the standard deviation is computed after applying the same reranker to the candidate sets produced by the three retriever runs.

F.2 Additional English BEL Evaluation on MedMentions

Table 8 reports R@1 on MedMentions, which we use as an English BEL transfer check. Unlike the fully supervised MedMentions baselines, BioELX does not use the MedMentions training set. In this annotation-free setting, our BioELX retriever improves over $\text{SapBERT}_{\text{multi}}$ from 50.3 to 53.3 (+3.0), and the full retrieve–rerank system reaches 60.1. Fully supervised systems remain stronger because they are trained end-to-end on MedMentions annotations, so this comparison is not supervision-matched. Nevertheless, the gains over $\text{SapBERT}_{\text{multi}}$ indicate that Wikidata-derived alias supervision and mention-anchored reranking also benefit English BEL without task-specific MedMentions training.

G Analysis of Retrieval Improvements

G.1 Do the retrieval gains come from cross-lingual alias bridges?

To better understand why the BioELX retriever improves over $\text{SapBERT}_{\text{multi}}$, we analyze the 1,416 improved mentions on XL-BEL, where our

Method↓	MedMentions (EN)
SapBERT$_{\text{multi}}$ (Liu et al., 2021b)	50.3✳
GenBioEL (Yuan et al., 2022b)	70.7
ArboEL (Agarwal et al., 2022)	**74.7**
Prompt-BioEL (Xu et al., 2023)	72.6
BioPro (Zhu et al., 2024)	66.5
GenBioEL+ANGEL (Kim et al., 2025)	<u>73.3</u>
BioELX retriever (**ours**)	53.3
BioELX retriever+reranker (**ours**)	60.1

Table 8: R@1 (%) on MedMentions. ✳ indicates reproduced results. Best result is bold, second best is underlined. All listed methods are trained on MedMentions except for ours and SapBERT$_{\text{multi}}$. For BioELX, we report the mean over three random seeds; standard deviations are provided in Appendix F.1.

Lang.	Imp. (%)	Group A	Group B	Group C
DE	152 (15.2)	0	127 (83.6)	25 (16.4)
EN	68 (6.8)	0	59 (86.8)	9 (13.2)
ES	74 (7.4)	0	57 (77.0)	17 (23.0)
FI	143 (14.3)	0	111 (77.6)	32 (22.4)
JA	118 (11.8)	0	99 (83.9)	19 (16.1)
KO	197 (19.7)	0	150 (76.1)	47 (23.9)
RU	152 (15.2)	0	120 (78.9)	32 (21.1)
TH	204 (20.4)	0	186 (91.2)	18 (8.8)
TR	121 (12.1)	0	107 (88.4)	14 (11.6)
ZH	187 (18.7)	0	150 (80.2)	37 (19.8)
All	1416 (14.2)	0	1166 (82.3)	250 (17.7)

Table 9: Group composition of improved mentions on XL-BEL. **Lang.** denotes the language code. **Imp. (%)** denotes the number of improved mentions and their percentage among test mentions in that language. Group A: only same-language Wikidata coverage; Group B: Wikidata coverage in at least one other language; Group C: no Wikidata coverage. Percentages in each group are computed over improved mentions.

retriever yields the largest gains. We define an *improved mention* as a test mention for which SapBERT$_{\text{multi}}$ fails to rank the gold CUI at top-1, while the BioELX retriever ranks the gold CUI at top-1.

We group these improved mentions according to the Wikidata-derived alias coverage of their gold CUI after decontamination. Group A contains mentions whose gold CUI has only same-language Wikidata aliases. Group B contains mentions whose gold CUI has aliases in at least one other language, providing potential cross-lingual alignment bridges. Group C contains mentions whose gold CUI has no Wikidata-derived alias coverage. Table 9 reports the group composition for each language.

The results show that the improvements are not driven by same-language alias coverage alone: Group A is empty for all languages. Instead, 82.3% of improved mentions belong to Group B, where the gold CUI is connected to aliases in other languages. This supports our cross-lingual alias bridge interpretation: multilingual aliases can anchor the correct concept across languages and help unseen non-English mentions retrieve the correct CUI, even when direct same-language coverage is limited. The remaining 17.7% of improvements fall into Group C, suggesting that continued training also improves the global embedding space beyond concepts directly covered by Wikidata aliases, although most improved mentions are associated with cross-lingual alias bridges.

We extend the analysis to the full XL-BEL dataset. Table 10 groups all 10,000 XL-BEL mentions by the Wikidata-derived alias coverage of their gold CUIs. The *No coverage* bin contains mentions whose gold CUI has no Wikidata-derived aliases in the retriever training data. The *Same-lang only* bin contains mentions whose gold CUI has aliases only in the mention language. The remaining bins correspond to gold CUIs with increasing amounts of other-language alias coverage.

BioELX improves over SapBERT$_{\text{multi}}$ in every non-empty coverage bin. The gain in the *No coverage* bin (+9.3 R@1) suggests that Wikidata-derived training does not only benefit CUIs directly covered by the added aliases. Instead, continued contrastive training may also improve the global geometry of the embedding space, reducing the model's reliance on superficial lexical similarity and improving generalization to uncovered concepts. At the same time, the largest gain appears in the 21+ bin (+12.3 R@1), where gold CUIs have broad other-language alias coverage. This supports the cross-lingual bridge interpretation: richer multilingual alias bridges provide stronger concept-level supervision. Together with Table 9, these results suggest that the retrieval gains are not explained by same-language alias memorization, but are largely associated with cross-lingual alias bridges and improved multilingual representation learning.

G.2 Dataset Ambiguity and Retriever Gains

To understand why the retriever gains are largest on XL-BEL, we compare the degree of surface-form ambiguity across benchmarks. We define a mention as ambiguous if the same mention string appears in the benchmark with more than one gold CUI. This

Alias coverage	Mentions	% test mentions	SapBERT$_{\text{multi}}$ R@1	BioELX R@1	Δ
No coverage	2,263	22.6	30.4	39.7	+9.3
Same-lang only	0	0.0	–	–	–
2–5 other langs	182	1.8	80.8	84.1	+3.3
6–20 other langs	949	9.5	51.7	61.0	+9.3
21+ other langs	6,606	66.1	35.7	48.0	+12.3

Table 10: XL-BEL mentions grouped by the Wikidata-derived alias coverage of the gold CUI in the retriever training data. **% test mentions** denotes the percentage of XL-BEL test mentions in each coverage bin. *No coverage* means that the gold CUI has no Wikidata-derived aliases in the training data. *Same-lang only* means that aliases are available only in the mention language. The remaining bins indicate the number of languages different from the mention language in which the gold CUI has aliases. Results are R@1 (%). Δ denotes the absolute improvement of BioELX over SapBERT.

statistic measures cases where the mention string alone does not uniquely determine the target entity, and where context is more likely to be needed for disambiguation.

Dataset	Mentions	Ambiguous	%
XL-BEL	10,000	69	0.7
WikiMed-DE	41,120	2,022	4.9
Patent	1,043	174	16.7
EMEA	2,155	665	30.9

Table 11: Surface-form ambiguity across benchmarks. We count a mention as ambiguous if its mention string appears in the benchmark with more than one gold CUI.

Table 11 shows that XL-BEL has the lowest surface-form ambiguity: only 0.7% of its mentions share the same surface form with another entity linked to a different CUI. WikiMed-DE is also relatively low at 4.9%. In contrast, EMEA and Patent are substantially more ambiguous, with 30.9% and 16.7% ambiguous mentions, respectively.

These differences help explain retrieval gains in the main experiments. Since our retriever is trained from aliases, it primarily improves alias-level matching. This is especially effective for benchmarks with low surface-form ambiguity, where the mention string itself provides strong evidence for the target entity. In more ambiguous benchmarks such as EMEA and Patent, the same mention string more often maps to different CUIs, so improving alias matching alone is less sufficient, and context-aware reranking becomes more important. This also helps explain why EMEA and Patent benefit more from context-aware reranking than from improvements in retrieval alone.

H LLM Reranker Adaptation Analyses

H.1 Mention-Anchored Prompting versus Lightweight Supervised Fine-Tuning

As discussed in Section 3.4, off-the-shelf LLM rerankers can be distracted by salient context tokens rather than the target mention span, leading to incorrect linking decisions and limiting their direct applicability to entity linking. Recent work also highlights the value of training-free mechanisms for focusing LLMs on task-relevant evidence (Zhong et al., 2026). For EL, the corresponding challenge is mention-level focus: rerankers must be adapted to score candidates with respect to the target mention rather than the context as a whole. We examine whether this adaptation requires supervised parameter updates or whether it can be achieved through a training-free input refinement. We conduct this ablation with QWEN3-Reranker, the main reranker used in our main paper, and compare two adaptation strategies: (i) *refining* the input with an explicit mention-anchored prompt, as shown in Figure 2, and (ii) *fine-tuning* the reranker on a small Wikipedia–Wikidata dataset to better follow the entity-linking instruction. Table 12 compares three variants:

1. Vanilla QWEN with its default template (*Orig. prompt*)
2. Vanilla QWEN with our mention-anchored prompt (*Refined prompt*), and
3. QWEN further fine-tuned using our supervision while keeping the same mention-anchored prompt (*Refined prompt + FT*).

For fine-tuning, we construct a lightweight dataset from Wikipedia and Wikidata with 1000 instances. Each instance contains one positive entity

	XL-BEL	EMEA	Patent
Orig. prompt	45.6	35.8	54.4
Refined prompt	**54.8**	**63.2**	**74.0**
Refined prompt + FT	54.0	61.5	71.6

Table 12: Ablation of prompt design vs. fine-tuning for the QWEN3-Reranker. All variants rerank the same top-k candidate sets produced by the same retriever (fixed k), and differ only in the input prompt/template and whether supervised fine-tuning is applied. Metrics are R@1 (%).

alias and 32 negative entity aliases. We fine-tune the reranker with parameter-efficient adaptation using LoRA (Hu et al., 2022) with rank $r = 16$, a learning rate of 1×10^{-4}, and one training epoch. We vary the language used for fine-tuning by creating three 1000-instance subsets: German-only, Thai-only, and a mixture across all XL-BEL languages. Among the fine-tuned variants, German performs best, but it still underperforms the non-fine-tuned baseline. To test whether the large number of negatives harms learning, we reduce the number of negatives to 7, but the results remain largely unchanged. We also try alternative learning-rate schedules, including smaller learning rates, without improving over the baseline.

As shown in Table 12, the mention-anchored prompt yields the strongest performance, while additional fine-tuning degrades the results of vanilla QWEN with our mention-anchored prompt. We attribute this to two factors. First, due to limited compute, we use LoRA instead of full-model fine-tuning: its restricted update capacity may be insufficient to induce the behavioral changes required for entity linking. Second, the fine-tuning instances are constructed from retrieval outputs. This dataset is small and may inherit retriever-specific biases, creating a distribution mismatch with evaluation data and limiting generalization. We expect that fine-tuning could become beneficial with a larger and more diverse supervision set and with reduced retrieval-induced bias; improving data selection and coverage is an important direction for future work.

H.2 Does Mention-Anchored Prompting Transfer Across LLM Rerankers?

Having shown that mention-anchored prompting is more effective than lightweight fine-tuning for QWEN3-Reranker, we next examine whether this adaptation transfers beyond a single reranker. Table 13 evaluates whether mention-anchored prompting improves reranking across different reranker families and model sizes. We conduct this analysis on EMEA and Patent because they are challenging benchmarks for alias-based retrieval. The same mention string can correspond to different gold entities depending on context, making them suitable for testing context-aware disambiguation. For example, in EMEA, the mention *Chemotherapie* is annotated with two different gold CUIs, C0392920 (Chemotherapy Regimen) and C0013216 (Pharmacotherapy), depending on the surrounding text.

We compare the original query–document formulation with our mention-anchored input for Jina-Reranker-V3[9], BGE-Reranker[10], and QWEN3-Reranker[11]. Our mention anchoring improves all rerankers on both benchmarks. The average gains are +7.09 R@1 on EMEA and +8.70 on Patent for Jina, +3.60 and +4.21 for BGE, +3.12 and +8.40 for QWEN3-Reranker-0.6B, +12.50 and +19.89 for QWEN3-Reranker-4B, and +18.58 and +22.02 for QWEN3-Reranker-8B, respectively.

These results suggest that mention-anchored prompting is broadly useful for adapting rerankers trained for general query–document relevance ranking to context-aware entity linking. At the same time, model capacity and the ranking interface both affect the magnitude of the gain. The improvement is largest for QWEN3-Reranker-8B, which combines larger model capacity with an instruction interface that allows us to explicitly steer the reranker to score candidates with respect to the marked target mention rather than generic query–document relevance. We therefore use QWEN3-Reranker-8B as the main reranker in our primary results, while using the other rerankers as additional checks of whether mention anchoring transfers across ranking architectures.

Since these rerankers are used as frozen pretrained models and no task-specific training is performed, this ablation does not involve training random seeds.

H.3 Mention Marker Ablation

Having shown that mention anchoring improves reranking across different rerankers, we further

[9] jina-reranker-v3: https://huggingface.co/jinaai/jina-reranker-v3

[10] bge-reranker-v2-gemma: https://huggingface.co/BAAI/bge-reranker-v2-gemma

[11] Qwen3-Reranker-8B: https://huggingface.co/Qwen/Qwen3-Reranker-8B

Dataset→			EMEA						Patent			
Reranker↓	Model size	**Input↓**	ES	FR	NL	DE	**Avg.**	**Gain**	FR	DE	**Avg.**	**Gain**
jina-reranker-v3	0.6B	Original	47.92	47.33	45.62	44.94	46.45	–	51.46	47.41	49.44	–
		Ours	55.56	54.76	53.23	50.59	53.54	+7.09	60.53	55.75	58.14	+8.70
Qwen3-Reranker-0.6B	0.6B	Original	44.21	43.85	45.62	40.55	43.56	–	40.06	38.22	39.14	–
		Ours	46.52	46.00	48.31	45.88	46.68	+3.12	47.66	47.41	47.54	+8.40
bge-reranker-v2-gemma	3B	Original	54.63	49.88	50.23	49.41	51.04	–	51.17	54.31	52.74	–
		Ours	57.41	53.60	54.61	52.94	54.64	+3.60	55.85	58.05	56.95	+4.21
Qwen3-Reranker-4B	4B	Original	45.60	48.49	44.01	44.47	45.64	–	46.49	49.71	48.10	–
		Ours	58.80	57.54	57.14	59.06	58.14	+12.50	69.30	66.67	67.99	+19.89
Qwen3-Reranker-8B	8B	Original	49.54	44.32	42.17	42.59	44.66	–	54.97	47.70	51.34	–
		Ours	**67.59**	**63.11**	**60.37**	**61.88**	**63.24**	**+18.58**	**75.73**	**70.98**	**73.36**	**+22.02**

Table 13: Effect of mention anchoring across different rerankers. All models rerank the same top-k candidate sets. For each reranker, the mention-anchored input explicitly marks the target mention span in the query. Results are R@1 (%). Avg. is computed separately for each benchmark, and Gain denotes the improvement of mention-anchored input over the original input.

examine which marker should be used to indicate the target span. Table 14 compares several mention-marking strategies while keeping the retriever, candidate set, and reranker fixed. All explicit markers improve over the no-marker setting, confirming that directing the reranker to the target span is important for the entity linking task. The `<tgt>...</tgt>` marker performs best across all EMEA and Patent settings. We attribute this gain to the fact that `<tgt>...</tgt>` follows a familiar span-highlighting convention (similar to HTML-style delimiters), making it easier for the reranker to reliably focus on the target mention when comparing candidates.

I Additional Ablations

I.1 How Does the Retriever Affect Reranking?

Since our reranker only operates on the top-k retrieved candidates, the retriever determines an upper bound on the achievable reranking accuracy: if the gold entity is not contained in $C_k(m)$, no reranking model can recover it. To quantify the dependency of the reranker on the retriever, we plug candidate sets produced by different retrievers into the same QWEN3-Reranker and compare the final R@1. We consider a range of retrievers with varying strengths, including BM25 (Robertson and Zaragoza, 2009) (sparse lexical retrieval), mBART (a weaker neural baseline), BERGAMOT (a strong cross-lingual system), and our BioELX retriever.

Table 15 reports each retriever's R@1 and R@k (with k=64) together with the reranked R@1 using identical reranking settings on XL-BEL. The results show a clear trend: stronger retrievers yield substantially better reranked performance, largely because they place the gold entity in the top-k set more often. For example, mBART has relatively low recall at k=64 (26.2), whereas BERGAMOT improves this to 46.6, and our retriever further increases it to 70.7; correspondingly, reranking accuracy improves as candidate coverage increases. These findings highlight that improving candidate retrieval is a prerequisite for effective contextual reranking in large biomedical knowledge bases.

I.2 Does Entity-Side Metadata Help Reranking?

Except for entity aliases, the knowledge base also provides entity metadata, including type and description. We examine whether providing additional entity metadata can improve LLM reranking beyond using the entity alias alone. Specifically, we compare three input settings for the reranker: (i) entity alias only (A), (ii) alias plus semantic type (A+T), and (iii) alias plus type and textual description (A+T+D).

Entity-side input to reranker	R@1
Alias only (A)	54.8
Alias + Semantic Type (A+T)	53.5
Alias + Type + Description (A+T+D)	52.5

Table 16: Effect of adding entity-side metadata to the QWEN3-Reranker on XL-BEL. All variants rerank the same top-k candidate sets produced by the same retriever (fixed k).

Mention marker	EMEA				Patent	
	ES	FR	NL	DE	FR	DE
No explicit marker	62.70	58.47	53.23	56.71	67.54	60.92
...	63.89	61.95	57.83	59.76	71.93	64.37
<mention>...</mention>	65.97	62.18	58.99	60.12	73.39	68.97
<tgt>...</tgt>	**67.59**	**63.11**	**60.37**	**61.88**	**75.73**	**70.98**

Table 14: Effect of different mention markers for reranking. All settings rerank the same top-k candidate sets and differ only in how the target mention is marked in the query. Results are R@1 (%). Test reranker is QWEN3-Reranker.

Retriever	R@1	R@64	+ Reranker R@1
BM25	19.7	26.9	26.1
mBART	17.2	26.2	25.9
BERGAMOT	35.6	46.6	37.3
BioELX retriever	50.1	70.7	**54.8**

Table 15: Effect of retriever quality on reranking. The reranker is fixed (QWEN3-Reranker with the same prompt and k=64). R@k provides an upper bound on reranked R@1. We report average results over the XL-BEL benchmark.

As shown in Table 16, adding entity types or descriptions does not improve performance and even slightly degrades R@1. We attribute this to two factors. First, within the top-k retrieved candidates, many entities already share the same (or highly similar) semantic types, so type information provides little discriminative signal but adds extra tokens that can distract the reranker. Second, entity descriptions have limited coverage for the entities appearing in our evaluation: in XL-BEL, only a small fraction of entities provide non-empty descriptions (roughly 10%), making the A+T+D setting inconsistent across candidates and often uninformative.

J Use of AI Assistants

We used AI assistants only for writing polishing, such as grammar correction and minor rephrasing.

K Additional Figures

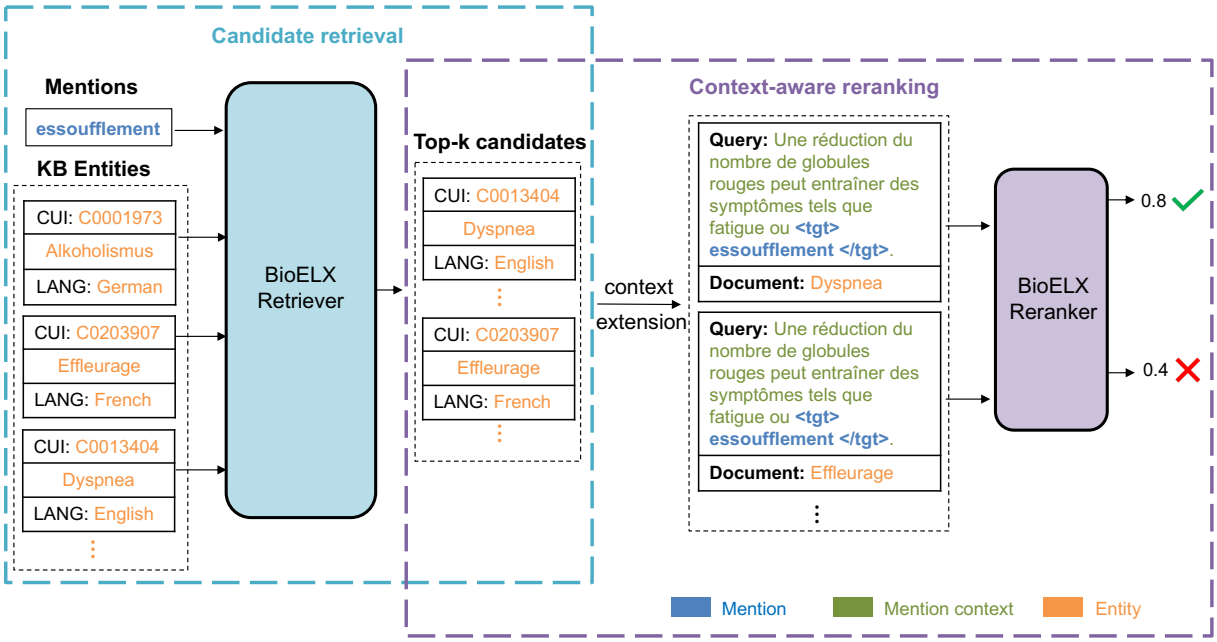

Figure 4: The two-stage BioELX pipeline, illustrated for a single mention. The retriever embeds the mention and multilingual entity aliases into a shared space, and returns the top-k aliases with associated entities by cosine similarity. For each candidate alias, the reranker scores a *mention-anchored query* — the mention and its surrounding context — against a document built from that alias, and selects the entity associated with the highest-scoring alias.

<|im_start|>system
You are a multilingual biomedical entity linking reranker. Judge whether the Document meets the requirements based on the Query and the Instruction provided. Note that the answer can only be "yes" or "no". <|im_end|>
<|im_start|>user
<Instruction>: {Given a Query (in any language) containing a target mention marked by <tgt>...</tgt> and its local context (nearby words/sentences), and a Document containing a candidate entity name in any language, decide whether the candidate is the correct link for the mention in context.}
<Query>: {Une réduction du nombre de globules rouges peut entraîner des symptômes tels que fatigue ou **<tgt> essoufflement </tgt>**}
<Document>: {Dyspnea}<|im_end|>
<|im_start|>assistant
<think>\n\n</think>\n\n

Figure 5: Prompt format for the BioELX Reranker, instantiated with the QWEN3-Reranker template (black text). The **query** contains the biomedical text T with the target mention m marked by `<tgt>...</tgt>`, and the **document** contains an alias a of the candidate entity e, so the reranker scores each mention-candidate pair. Our "document" is intentionally short: many KB entities lack rich textual metadata, so we provide only the entity alias.

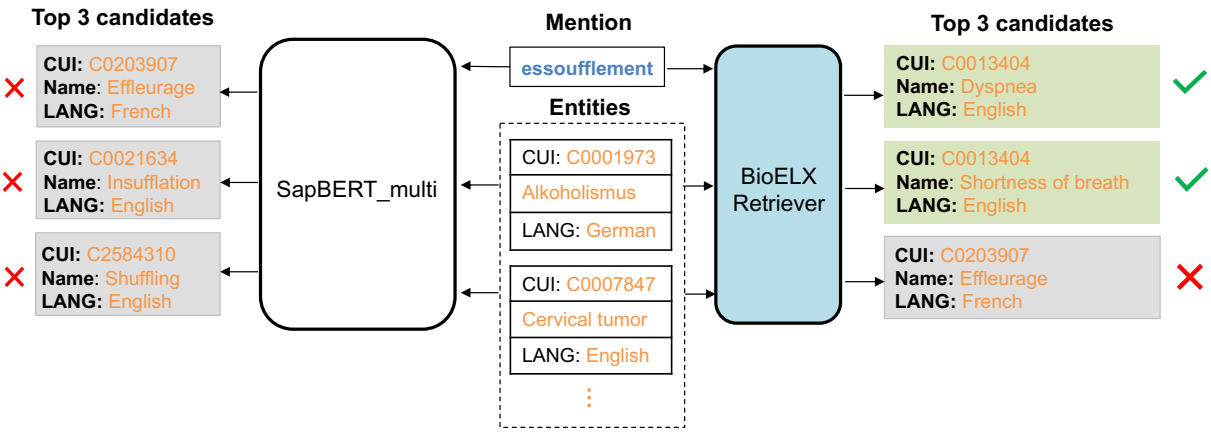

Figure 6: An example of retrieval. For the French mention *essoufflement*, SapBERT retrieves top candidates based on surface similarity but misses the correct concept. Our retriever, enhanced with Wikidata-derived multilingual aliases, retrieves candidates that include the correct CUI (Dyspnea) among the top results.