Proceedings of the Thirty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-26)
Special Track on AI and Health

SynCABEL: Synthetic Contextualized Augmentation for Biomedical Entity Linking

Adam Remaki[1], **Christel Gérardin**[1,2], **Eulàlia Farré-Maduell**[3], **Martin Krallinger**[3] and **Xavier Tannier**[1]

[1] Sorbonne Université, Inserm, Université Sorbonne Paris Nord, Limics, 75006 Paris, France
[2] Service de médecine interne, Hôpital Tenon, Assistance Publique - Hôpitaux de Paris, Paris, France
[3] Barcelona Supercomputing Center, Barcelona, Spain
adam.remaki@etu.sorbonne-universite.fr, christel.gerardin@aphp.fr,
{eulalia.farre,martin.krallinger}@bsc.es, xavier.tannier@sorbonne-universite.fr

Abstract

We present **SynCABEL** (Synthetic Contextualized Augmentation for Biomedical Entity Linking), a framework that addresses a central bottleneck in supervised biomedical entity linking (BEL): the scarcity of expert-annotated training data. SynCABEL leverages large language models to generate context-rich synthetic training examples for all candidate concepts in a target knowledge base, providing broad supervision without manual annotation. We demonstrate that SynCABEL, when combined with decoder-only models and guided inference, establishes new state-of-the-art results across three widely used multilingual benchmarks: MedMentions for English, QUAERO for French, and SPACCC for Spanish. Evaluating data efficiency, we show that SynCABEL reaches the performance of full human supervision using up to 60% less annotated data, substantially reducing reliance on labor-intensive and costly expert labeling. Finally, acknowledging that standard evaluation based on exact code matching often underestimates clinically valid predictions due to ontology redundancy, we introduce an LLM-as-a-judge protocol. This analysis reveals that SynCABEL significantly improves the rate of clinically valid predictions. Our synthetic datasets, models, and code are released to support reproducibility and future research:

- **HuggingFace Datasets & Models**
- **GitHub Repository**

1 Introduction

Biomedical Entity Linking (BEL) is the task of mapping spans of text in biomedical documents to concepts in knowledge bases (KBs) such as the UMLS or SNOMED CT.

Early approaches relied on context-free lexical matching or self-supervised models trained from the KB alone. However, they struggle with ambiguity: terms such as "discharge" (release vs. secretion) or abbreviations such as "AS" (Aortic Stenosis vs. Ankylosing Spondylitis) may refer to distinct concepts depending on context [Newman-Griffis *et al.*, 2021; Xu *et al.*, 2007]. These limitations motivated context-aware BEL models, which encode mentions with surrounding text to disambiguate surface forms. Although effective when sufficiently trained, they introduce a major constraint: the need for large-scale, high-quality annotated datasets, which are scarce in the biomedical domain. Creating such resources is labor-intensive, requiring clinical experts to match mentions to ontology concepts. This scarcity limits generalization and creates a bottleneck for supervised BEL, as illustrated in Figure 1.

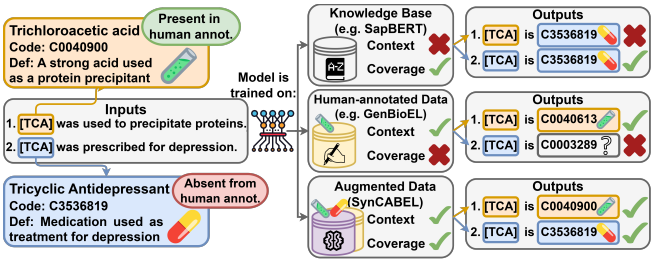

Figure 1: Illustration of biomedical entity linking annotation scarcity: a context-free model fails without context, a supervised context-aware model trained on human annotations fails on unseen concepts, while a SynCABEL-augmented model leverages synthetic data to recover the correct concepts.

To address this annotation scarcity issue, we introduce **SynCABEL** (**Syn**thetic **C**ontextualized **A**ugmentation for **B**iomedical **E**ntity **L**inking), a novel framework designed to enhance context-aware BEL by leveraging large language models (LLMs) to generate context-rich training instances for all candidate concepts in a KB. Our contributions are listed as follows:

- We release the first large-scale multilingual synthetic dataset for BEL in English, French, and Spanish.
- We show that SynCABEL matches full human-annotated training performance with substantially fewer human annotations.
- We demonstrate that combining recent decoder-only models with SynCABEL and guided inference achieves state-of-the-art performance on multiple BEL benchmarks.

Additionally, we developed an LLM-as-a-judge evaluation protocol for BEL that goes beyond exact code matching by assessing the semantic relationship between predicted and gold concepts, including equivalence, broader, narrower, and unrelated relations.

2 Related Work

2.1 BEL Systems

Research in BEL has evolved through several distinct paradigms, each addressing the problem with different strengths and limitations [Chen *et al.*, 2026].

Rule-based Systems. Early systems like cTAKES [Savova *et al.*, 2010], and SciSpacy [Neumann *et al.*, 2019] relied on heuristic string matching. While widely adopted, their reliance on static dictionaries limits generalization to predefined lexical variants, causing failures when a mention refers to an existing concept using a surface form not listed among its synonyms in the ontology.

Context-free Bi-encoder Models. Transformer-based models like SapBERT [Liu *et al.*, 2021a; Liu *et al.*, 2021b] and CODER [Yuan *et al.*, 2022b] use self-supervised contrastive learning on synonyms. While efficient for embedding names in a shared space without annotated data, they ignore surrounding context, limiting their ability to resolve ambiguity.

Prompting LLMs. LLMs have been used via prompting to simplify mentions before linking [Borchert *et al.*, 2024; Vollmers *et al.*, 2025], as re-rankers [Ye and Mitchell, 2025], or guided via constrained decoding without training [Lin *et al.*, 2025]. However, prompting-based methods often yield inferior results compared to specialized EL models.

Contextualized Bi-encoder Models. These systems encode mentions with context and candidates in a shared space. Unlike context-free models, they explicitly learn to capture contextual information. ArboEL [Agarwal *et al.*, 2022], for example, adapts BLINK [Wu *et al.*, 2020] and introduces a graphical arborescence objective to model cross-document coreference through directed spanning trees, achieving state-of-the-art results on MedMentions [Mohan and Li, 2018].

Generative Models. These models treat entity linking as conditional text generation, directly producing concept identifiers. GENRE [Cao *et al.*, 2021] introduced constrained decoding to ensure valid outputs and strictly improve memory efficiency by avoiding dense indexes. Currently, biomedical implementations like GenBioEL [Yuan *et al.*, 2022a] are built on pretrained encoder-decoder architectures. The general-domain InsGenEL [Xiao *et al.*, 2023] has extended this approach by fine-tuning a decoder-only model, though it is optimized for a joint mention detection and linking objective and not for entity linking only.

2.2 BEL Data Augmentation

Despite architectural differences, both retrieval and generative methods depend on annotated datasets that currently cover only a fraction of the KB and lack variability. This scarcity creates a bottleneck for developing robust, generalizable BEL systems, motivating alternative data creation strategies to improve scalability and coverage while reducing annotation costs.

Weak and Distant Supervision. Approaches using exact string matching in large corpora (e.g., PubMed, Wikipedia) to create training data [Zhang *et al.*, 2022; Vashishth *et al.*, 2021; Wang *et al.*, 2023] are scalable but noisy due to synonym ambiguity and non-clinical contexts.

Template-Based Synthetic Data. To better capture entity semantics, Yuan *et al.* [2022a] used UMLS definitions as templates, filled with synonyms. This provides clean, concept-focused data but lacks the lexical and contextual diversity of natural language.

LLM-generated Synthetic Data. Xin *et al.* [2025] and Chen *et al.* [2025] have demonstrated the effectiveness of LLMs for generating additional training examples, but restrict synthesis to concepts already observed during training. While this increases contextual diversity and improves generalization across known entities, it does not address annotation scarcity for entirely unseen concepts. Relatedly, Josifoski *et al.* [2023] leverage synthetic data to mitigate data scarcity, but their approach targets a different task setting, namely relational triple extraction.

2.3 Positioning of SynCABEL

SynCABEL advances BEL on two fronts. First, it introduces a data augmentation strategy that generates examples for the entire space of candidate concepts in the KB. It overcomes the limitations of prior work that either used templates for full coverage (low quality) or LLMs only for training-set concepts (limited coverage). Second, it is the first approach to fine-tune a decoder-only model specifically for BEL, allowing us to leverage recent improvements in large foundation models.

3 Methodology

3.1 Problem Statement

Let $\mathcal{E}$ represent the set of candidate concepts from a given KB. Each concept $e \in \mathcal{E}$ is associated with a semantic group g_e (e.g., *Disorders*) and multiple synonyms $S_e = \{s_e^1, \ldots, s_e^{n_e}\}$. Given a mention m with left context c_l, right context c_r and semantic group g_e, our objective is to predict the correct concept e_m that the mention m refers to.

3.2 Generative BEL Training Objective

We adopt the autoregressive formulation from Cao *et al.* [2021]. The input sequence x is defined as:

$$x = c_l \; [\; m \;] \; \{\; g_e \;\} \; c_r \; [\; SEP \;] \; [\; m \;] \text{ is}$$

where c_l, c_r denote context, m the target mention, and g_e its semantic group. The token $[SEP]$ and cue phrase "m is" prompt the model to generate output sequence y defined as:

$$y = e_m$$

where e_m is the concept corresponding to mention m. We maximize the conditional likelihood:

$$p_\theta(y|x) = \prod_{i=1}^{N_y} p_\theta(y_i|y_{<i}, x)$$

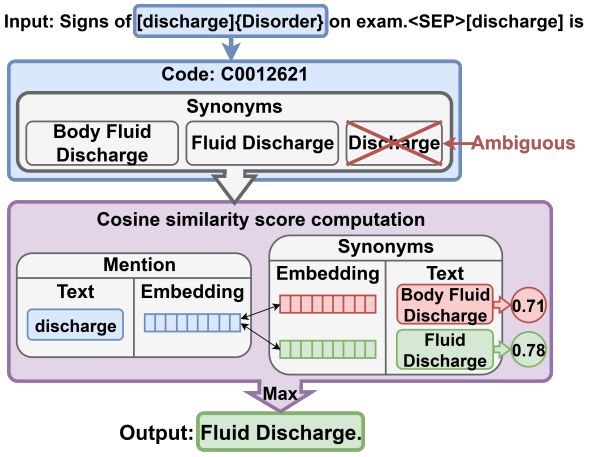

Figure 2: Example of the disambiguation process used to select a natural representation e_m = "Fluid Discharge" for a concept e = "C0012621" and a mention m = "discharge".

where N_y is the number of tokens in the output, and y_i is the i-th token.

3.3 Adaptive Concept Representation for Training

A key challenge during training is representing the concept e_m in natural language for the target output. We employ an adaptive method that selects the most appropriate synonym of a concept based on the input mention. As shown in Figure 2, the method proceeds in two steps. First, we preprocess each concept's list of synonyms by removing those that are ambiguous with other concepts in the same semantic group. For example, the term "discharge" is removed from the *Disorder* group because it can refer to multiple clinical symptoms. Then, for each remaining unambiguous synonym s_e, we compute the cosine similarity between the embeddings of the mention m and s_e. The concept representation e_m is defined as the synonym with the highest similarity score.

$$e_m = \arg\max_{s_e \in \mathcal{S}_e} \frac{\mathbf{v}_m \cdot \mathbf{v}_{s_e}}{\|\mathbf{v}_m\| \|\mathbf{v}_{s_e}\|}$$

where $\mathbf{v}_m$ and $\mathbf{v}_{s_e}$ are the embedding vectors of the mention m and synonym s_e, respectively. Importantly, the embeddings can be derived from any model, in our experiments, we compare different embedding models to assess their impact on linking performance.

3.4 Synthetic Data Generation

To improve concept coverage while maintaining contextual realism, we generate synthetic training data using a structured prompt-based approach. For a given concept e, our method generates contextual sentences that capture diverse, natural ways of expressing e. As illustrated in Figure 3, the prompt contains three components: (i) a task description that decomposes generation into two steps, first generating mentions of the concept and then generating contextualized sentences for each mention. (ii) Random contextual examples from the human-annotated training set. (iii) The title, semantic group, semantic type, definitions, and synonyms of the target concept from the KB. The prompt is written in the same language as the target dataset and provided to an LLM, which produces the final contextualized sentences.

3.5 Training Data Composition

The synthetic examples generated by SynCABEL are combined with human-annotated training data to fine-tune the entity linking model (Figure 3). To preserve the characteristics of clinical reports while benefiting from the increased coverage of synthetic data, we jointly train on both sources and upsample human-annotated examples such that they constitute half of the training instances.

3.6 Guided Inference

Greedy decoding often yields invalid entities absent from the KB. To address this, we adopt the guided inference mechanism from Cao *et al.* [2021], illustrated in Figure 3. First, the KB vocabulary is filtered by the target entity's semantic group (e.g., *Disorders*) and structured as a trie, where each path maps a synonym to a unique concept. Then, the model dynamically restricts token selection at each step to valid branches. For instance, after generating the prefix "A", the trie permits only valid continuations like "the" (for *Atherosclerosis*) or "ortic" (for *Aortic*).

3.7 LLM-as-a-judge Evaluation

Standard entity linking evaluation relies on exact code matching and therefore may fail to capture the true clinical relationship between predicted and gold concepts. In practice, a prediction can be clinically correct despite a code mismatch due to ontology redundancy (e.g., multiple codes for the same concept) or annotation noise, where the prediction better reflects the mention than the human label. To capture these nuances, we introduce a four-class scheme (Figure 4): **(i) Correct**, clinically appropriate even without exact code match; **(ii) Broad**, relevant but overly general; **(iii) Narrow**, relevant but overly specific; and **(iv) No relation**, clinically unrelated. Labels are assigned automatically by an LLM using precise class definitions and five clinician-annotated examples.

4 Experiments

4.1 Datasets

Human-Annotated Datasets. We rely on three human-annotated corpora for our experiments:

- **MedMentions-ST21pv (MM-ST21pv)** [Mohan and Li, 2018], a curated subset of the MedMentions corpus, consisting of 4,392 English PubMed abstracts. MedMentions is currently the largest publicly available human-annotated dataset for biomedical entity linking.
- **QUAERO French Medical Corpus (QUAERO)** [Névéol *et al.*, 2014], the largest French dataset for biomedical entity linking, composed of two subsets: QUAERO-MEDLINE, derived from MEDLINE titles, and QUAERO-EMEA, extracted from documents published by the European Medicines Agency.
- **Spanish Clinical Case Corpus (SPACCC)** [Miranda-Escalada *et al.*, 2022; Lima-López *et al.*, 2023a; Lima-López *et al.*, 2023b], a manually annotated collection of

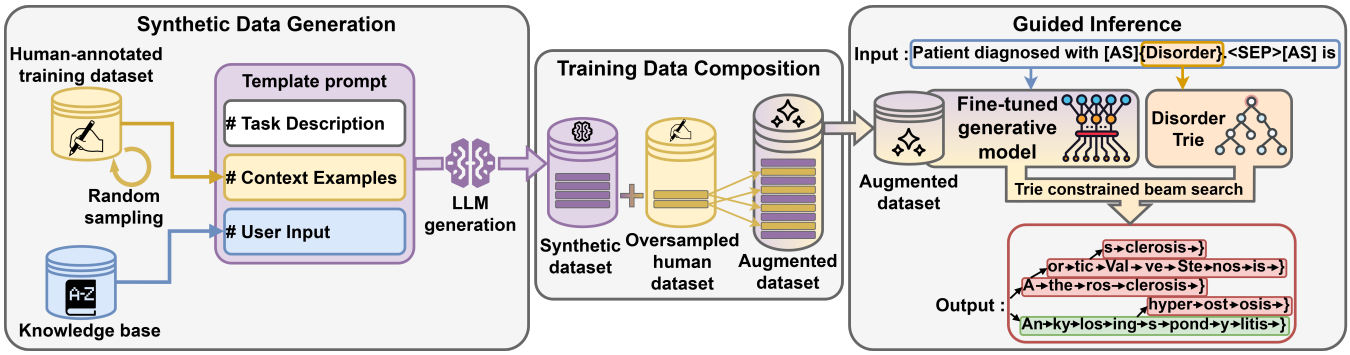

Figure 3: Overview of SynCABEL. LLM: large language model

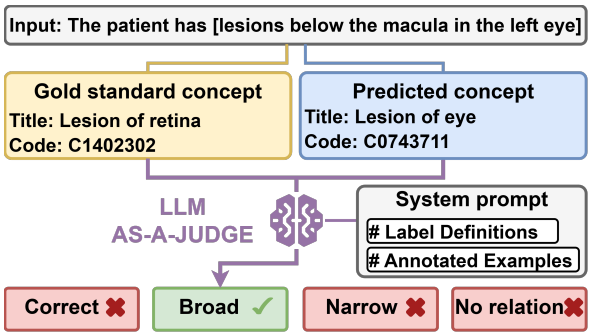

Figure 4: LLM-as-a-judge evaluation of predicted vs. gold concepts using four clinical-relation classes: Correct, Broad, Narrow, and No relation.

	MM-ST21PV	QUAERO	SPACCC
# Docs	4,392	2,422	1,000
# Mentions	203,282	16,185	27,799
# Concepts	25,419	5,045	9,254
# Sem. Groups	14	10	3
Language	English	French	Spanish
Clinical cases	✗	✗	✓

Table 1: Summary statistics of the human-annotated datasets. MM-ST21pv: MedMentions-ST21pv; SPACCC: Spanish Clinical Case Corpus

clinical case reports derived from open-access Spanish medical publications. The corpus contains 1,000 clinical cases, with disease, procedure, and symptom mentions annotated using the SNOMED CT ontology.

MM-ST21pv and QUAERO were selected because they are the largest human-annotated biomedical entity linking datasets available in their respective languages and because they cover nearly all UMLS semantic groups. SPACCC was chosen as the largest Spanish BEL dataset and, more importantly, because it consists of clinical reports, making it particularly valuable for assessing the potential applicability of our approach in clinical settings. Table 1 summarizes the main characteristics of these datasets.

Dataset	SynthMM	SynthQUAERO	SynthSPACCC
# Mentions	460,878	396,914	1,813,463
# Concepts	153,374	132,089	276,649

Table 2: Summary of synthetic datasets.

Synthetic Datasets. We constructed three synthetic datasets (SynthMM, SynthQUAERO, and SynthSPACCC) by using Llama-3-70B [Grattafiori *et al.*, 2024] to generate three contextualized training examples per concept, prompted with five random annotated examples from the target training set. For instance, SynthMM includes the concept *Apophysitis* (CUI C0264110) in the sentence: "The adolescent athlete presented with knee pain and swelling, diagnosed with [apophysitis] of the tibial tubercle, highlighting the importance of proper training and warm-up routines." While applicable to all candidate concepts, SynthMM and SynthQUAERO were restricted to UMLS concepts with definitions (6.5% and 4.5% of the KB, respectively) for computational feasibility. Table 2 summarizes the generated datasets. The full datasets, examples, and prompts are available on HuggingFace.

4.2 Implementation details

Hardware and Software. All experiments were conducted on a single NVIDIA H100 GPU (80GB). For reproducibility, full implementation details, hyperparameters, and code are provided on GitHub.

Knowledge Bases. We used three KBs aligned with each dataset's annotation guidelines. For MM-ST21pv, we used UMLS 2017AA restricted to the ST21-pv scheme (21 semantic types and synonyms from 18 source ontologies). For QUAERO, we used UMLS 2014AA filtered to 10 semantic groups. For SPACCC, we relied on gazetteers derived from relevant branches of Spanish SNOMED-CT (July 31, 2021 release); since SPACCC comprises three subsets (diseases, symptoms, procedures), these categories served as semantic groups.

Model	MM-ST21PV (english)	QUAERO-MEDLINE (french)	QUAERO-EMEA (french)	SPACCC (spanish)	Average
RULE-BASED (UNSUPERVISED)					
SciSpacy [Neumann *et al.*, 2019]	53.8	40.5	37.1	13.2	36.2
CONTEXT-FREE BI-ENCODER (SELF-SUPERVISED)					
SapBERT [Liu *et al.*, 2021a]	51.1	50.6	49.8	33.9	46.4
CODER-all [Yuan *et al.*, 2022b]	56.6	58.7	58.1	43.7	54.3
SapBERT-all [Liu *et al.*, 2021b]	64.6	74.7	67.9	47.9	63.8
CONTEXTUALIZED BI-ENCODER (SUPERVISED)					
ArboEL [Agarwal *et al.*, 2022]	<u>74.5</u>	70.9	62.8	49.0	64.2
GENERATIVE ENCODER-DECODER (SUPERVISED)					
mBART-large [Tang *et al.*, 2020]	65.5	61.5	58.6	57.7	60.8
+ Guided inference [Cao *et al.*, 2021]	70.0	72.8	71.1	61.8	68.9
+ SynCABEL (Our method)	71.5	77.1	<u>75.3</u>	64.0	72.0
GENERATIVE DECODER-ONLY (SUPERVISED)					
Llama-3-8B [Grattafiori *et al.*, 2024]	69.0	66.4	65.5	59.9	65.2
+ Guided inference [Cao *et al.*, 2021]	74.4	<u>77.5</u>	72.9	<u>64.2</u>	<u>72.3</u>
+ SynCABEL (Our method)	**75.4**	**79.7**	**79.0**	**67.0**	**75.3**

Table 3: Entity linking performance (Recall@1) on biomedical benchmarks. The highlighted rows show our contribution, where the base model is trained on an augmented dataset. The best results are shown in **bold**, the second-best results are <u>underlined</u>, and the "Average" column reports the mean score across the four benchmarks.

4.3 Results

Benchmark Performance

We reproduced several state-of-the-art BEL systems from different paradigms: the rule-based **SciSpacy** [Neumann *et al.*, 2019]; the context-free bi-encoders **SapBERT** [Liu *et al.*, 2021a], **SapBERT-all** [Liu *et al.*, 2021b], and **CODER-all** [Yuan *et al.*, 2022b]; the contextualized bi-encoder **ArboEL** [Agarwal *et al.*, 2022]; the generative encoder-decoder **mBART-large** [Tang *et al.*, 2020] and decoder-only **Llama-3-8B** [Grattafiori *et al.*, 2024]. For generative models (mBART-large and Llama-3-8B), we evaluated three configurations: (i) supervised fine-tuning on human-annotated data; (ii) adding guided inference adapted from GENRE [Cao *et al.*, 2021]; and (iii) our proposed SynCABEL approach, which augments the training data with synthetic examples while retaining guided inference. For fair comparison, all baselines were constrained to the same KBs, and for each mention, only concepts from the matching semantic group were considered as candidates. In practice, this meant that for a given input, every baseline was provided with the same candidate set. This constraint could not be applied to SciSpacy, as its rule-based approach uses the entire KB and cannot be modified.

Table 3 presents entity linking performance (Recall@1) across four biomedical benchmarks: MM-ST21pv, QUAERO-MEDLINE, QUAERO-EMEA, and SPACCC. Models are grouped by architecture paradigm. Our SynCABEL framework with Llama-3-8B achieves the highest scores on all four benchmarks: 75.4 on MM-ST21pv, 79.7 on QUAERO-MEDLINE, 79.0 on QUAERO-EMEA, and 67.0 on SPACCC. The mBART-large variant of SynCABEL also shows consistent improvements over its baseline across all datasets.

	MM-ST21PV	QUAERO-MEDLINE	QUAERO-EMEA
Title (Static)	61.7	53.7	50.6
CODER-all	68.4	72.5	66.0
TF-IDF	**70.0**	**72.8**	**71.1**

Table 4: Recall@1 scores for different concept representation methods. mBART-large was supervised only on the human-annotated dataset (i.e., without data augmentation), and inference was performed using the guided inference method. The highlighted row shows the chosen representation method.

Adaptive Concept Representation for Training

We evaluate our adaptive concept representation strategy against a static baseline using concept preferred titles. We consider two adaptive representations: character-level 3-gram TF-IDF [Neumann *et al.*, 2019], following its effective use in GenBioEL [Yuan *et al.*, 2022a] and CODER-all embeddings [Yuan *et al.*, 2022b]. Experiments are conducted on UMLS-based datasets (MM-ST21pv, QUAERO-MEDLINE and QUAERO-EMEA) using mBART-large, with training limited to the original data (no augmentation) and guided decoding at inference.

As shown in Table 4, the TF-IDF-based representation con-

	MM-ST21pv	QUAERO-MEDLINE	QUAERO-EMEA	SPACCC
SPT	74.1	78.0	77.1	64.0
COMB	75.4	79.4	77.0	64.3
INT	**75.4**	**79.7**	**79.0**	**67.0**

Table 5: Recall@1 scores for different training data composition strategies. The highlighted row shows the chosen strategy.

sistently achieves the highest Recall@1 across all datasets, outperforming both the static baseline and CODER-all. We therefore adopt TF-IDF-based adaptive representations for all other experiments.

Training Data Composition

We evaluate three training data composition strategies when combining human-annotated and synthetic data: **(i) Synthetic Pretrain**, which uses a two-stage training procedure, first training on synthetic data and then fine-tuning on human data, without resampling; **(ii) Combined**, which performs a single-stage training on the union of human and synthetic datasets as-is, without resampling; **(iii) Interleaved**, which performs a single-stage training on merged data while oversampling human samples so that each training step alternates between a human and a synthetic instance. These experiments use Llama-3-8B with guided inference.
Recall@1 results in Table 5 show that interleaved oversampling consistently outperforms other strategies across all datasets, highlighting the importance of preserving a strong human signal while leveraging synthetic diversity. We therefore adopt this strategy in other experiments.

LLM-as-a-Judge Evaluation

To complement exact code matching, we employed GPT-5.2[1] as an automated judge to assess clinical validity on SPACCC, the only dataset containing full clinical reports. We evaluated all cases where the prediction of our best-performing model (Llama-3-8B) differed from the gold standard.[2]

A clinician labeled 150 mismatches across Disorders, Symptoms, and Procedures. GPT-5.2 achieved 66.2% agreement, with 79% precision for the 'Correct' label. Crucially, errors are conservative: only 12.4% of judge predictions overstated the human label (e.g., Narrow for No relation). This tendency to under-predict relevance implies the judge acts as a lower-bound estimator of validity.

Table 6 compares standard exact matching against the LLM-as-a-judge evaluation metric. While SynCABEL yields a 2.8% gain in exact matching, it achieves a larger 3.7% improvement in clinically valid predictions (rising from 68.9% to 72.6%). Given the judge's proven conservative bias, it shows SynCABEL helps the model capture correct semantics even when it misses the exact target code.

[1]Developed by OpenAI and accessed via their API.

[2]We also explored the KB-based architecture to explicitly detect broader/narrower relations, but it did not yield reliable results, likely due to the complexity and brittleness of these KBs.

Label	w/o SynCABEL	w/ SynCABEL
Exact Matching	64.2 (62.6–65.8)	67.0 (65.5–68.6)
Correct	**68.9 (67.4–70.4)**	**72.6 (71.2–74.1)**
Broad	13.0 (12.1–13.8)	11.0 (10.3–11.8)
Narrow	4.3 (3.8–4.7)	5.9 (5.3–6.5)
No relation	13.8 (11.1–16.7)	10.5 (7.6–13.2)

Table 6: Llama-3-8B performance on SPACCC with and without SynCABEL augmentation. Results are reported as percentages with 95% confidence intervals estimated via document-level empirical bootstrap, for exact code matching and each LLM-as-a-judge label.

5 Discussion

5.1 Annotation Scarcity

To assess synthetic data impact, we stratify performance by whether concepts appeared in training. As shown in Table 7, SynCABEL consistently improves unseen concepts performance across benchmarks while preserving seen results, with notable gains on SPACCC and QUAERO-EMEA (+9.4 and +9.9 points). However, unseen performance remains below seen (e.g., 30.2 vs 83.0 on SPACCC), showing that SynCABEL reduces but does not fully close the annotation scarcity gap. For UMLS-based datasets (QUAERO, MM), generation was restricted to concepts with definitions, leaving some "unseen" test concepts absent from augmented data, a filter that limits gains compared to SPACCC, suggesting that extending generation to all concepts could further bridge this gap.

Subset	w/o SynCABEL	w/ SynCABEL
MEDMENTIONS-ST21PV		
Overall	74.4 (73.6–75.4)	75.4 (74.5–76.3)
Seen Concepts	81.8 (81.0–82.6)	82.2 (81.4–83.0)
Unseen Concepts	46.0 (43.9–48.2)	49.0 (46.7–51.3)
QUAERO-MEDLINE		
Overall	77.5 (75.9–79.1)	79.7 (78.0–81.2)
Seen Concepts	90.5 (89.1–91.8)	89.7 (88.3–91.0)
Unseen Concepts	58.7 (55.8–61.6)	65.1 (62.1–68.1)
QUAERO-EMEA		
Overall	72.9 (68.3–77.3)	79.0 (75.7–82.2)
Seen Concepts	89.0 (86.1–91.6)	92.0 (89.6–94.3)
Unseen Concepts	52.5 (46.1–59.3)	62.4 (56.4–68.7)
SPACCC		
Overall	64.2 (62.6–65.8)	67.0 (65.5–68.6)
Seen Concepts	83.0 (81.8–84.2)	83.0 (81.8–84.2)
Unseen Concepts	20.8 (19.2–22.6)	30.2 (28.1–32.4)

Table 7: Exact-match Recall@1 performance on three biomedical benchmarks comparing two variants of Llama-3-8B: one trained only on the corresponding human-annotated dataset with guided inference, and one further enhanced with SynCABEL. Results are reported overall and for subsets defined by seen/unseen concepts. 95% confidence intervals were estimated via the empirical bootstrap method at the document level.

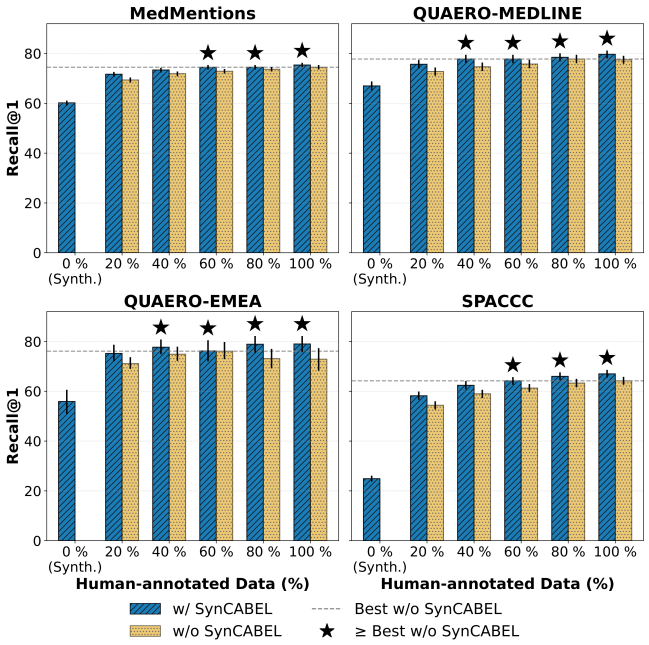

Figure 5: Data Efficiency Analysis. Exact-match Recall@1 performance as a function of human training data availability. The dashed line marks the performance ceiling of the standard baseline trained on human-annotated data. Stars (⋆) indicate where the SynCABEL-augmented model (blue) matches or surpasses this ceiling. 0% implies training on synthetic data only. Error bars denote 95% bootstrap CIs.

5.2 Reducing Annotation Effort

We evaluate data efficiency by varying the fraction of human-annotated training data (0–100%), with subsets created by randomly sampling documents, and comparing SynCABEL augmentation to the human-annotated only baseline.

Figure 5 demonstrates that SynCABEL substantially reduces annotation needs: it reaches full-data performance using 60% of human annotations on MM-ST21pv and SPACCC, and 40% on QUAERO. The gains are largest in low-resource regimes and diminish as more human data becomes available. This trend is clear on SPACCC and MM-ST21pv, though it is less observable on QUAERO due to the larger confidence intervals inherent to its smaller size. Finally, models trained solely on synthetic data significantly underperform fully supervised baselines, confirming that human annotations remain essential for optimal results on human-annotated data.

5.3 Real-World Applicability

BEL plays a central role in enabling the secondary use of electronic health records. By transforming unstructured free-text into structured concepts, it supports medical research, clinical decision, and downstream tasks such as automated clinical coding, cohort identification, and patient recruitment for clinical trials [Remaki *et al.*, 2025; French and McInnes, 2023]. To assess the viability of our models for such real-world deployment, we evaluate two dimensions: the utility of confidence scores for ensuring high-precision predictions, and the computational efficiency required for scale.

	Model (GB)	Cand. (GB)	Speed (/s)
SapBERT	2.1	20.1	**575.5**
ArboEL	**1.2**	7.1	38.9
mBART	2.3	**5.4**	51.0
Llama-3-8B	28.6	**5.4**	19.1

Table 8: Inference speed and memory footprint. Model: Model parameters size; Cand.: Candidate memory size.

Prediction Confidence. For each prediction, our model outputs a confidence score, computed as the mean token probability of the generated sequence. This score enables confidence-based filtering to trade precision for recall depending on application needs. On SPACCC, we empirically set a confidence threshold of 0.9 and evaluated performance on predictions exceeding this threshold.

It yields a precision of 80.8% (79.6–81.9), recall of 61.9% (60.2–63.5), and F1-score of 70.1% (68.5–71.6). These results illustrate how applying a strict confidence threshold can substantially boost precision for high-stakes applications while retaining reasonable recall.

Inference Efficiency and Memory Footprint. We evaluate real-world deployability based on throughput and memory usage (Table 8). While SapBERT achieves the highest speed, it requires a massive 20.1 GB candidate index. ArboEL uses a smaller BERT encoder, reducing model size (1.2 GB) and candidate memory (7.1 GB), but throughput is lower (38.9 mentions/s) due to the cross-encoder re-ranking step required at inference. In contrast, generative models rely on a compact candidate trie (5.4 GB), independent of model size. The mBART configuration processes 51.0 mentions/s with a small model footprint (2.3 GB), whereas Llama-3-8B requires substantially more GPU memory (28.6 GB) and achieves lower throughput (19.1 mentions/s).

6 Conclusion and Future Work

We introduced SynCABEL, a framework addressing annotation scarcity in BEL by generating synthetic, contextualized training examples for all candidate concepts in a target KB. By systematically enriching concept representations with diverse usage contexts, SynCABEL achieves state-of-the-art performance on multilingual benchmarks, improves recall on unseen concepts, and increases clinical validity under our LLM-as-a-judge protocol. While human annotations remain the strongest form of supervision, our results show that competitive performance can be achieved with substantially less of this costly data, thereby maximizing the value of expert effort.

Future directions include extending the generation context beyond single sentences to better capture document-level evidence, expanding to additional languages and domains, and refining training with negative sampling strategies such as ANGEL [Kim *et al.*, 2025]. We also aim to improve synthetic data quality through prompt refinement and expert intrinsic evaluation, and to reduce generation cost through smaller LLMs and selective concept sampling.

Ethical Statement

This work does not use private or identifiable patient records. Experiments rely on publicly available or licensed biomedical entity linking resources, together with synthetic data generated for research purposes. QUAERO was used under GFDL, MedMentions under CC0, and the UMLS Metathesaurus under its License Agreement, with acknowledgment of the National Library of Medicine. The released synthetic datasets are intended for research in biomedical entity linking and should not be used for clinical decision-making without appropriate validation.

Although synthetic data reduces privacy risks and dependence on costly expert annotation, it may still reflect biases, errors, or limitations of the source knowledge bases and generative models. We therefore encourage users to inspect the data before downstream use. Generating large-scale synthetic data also incurs computational and environmental costs; future work should balance data volume, model performance, and energy use. The authors used GitHub Copilot for code completion; all code was reviewed by the authors, who take full responsibility for it.

Acknowledgments

This work was performed using HPC resources from GENCI-IDRIS (Grant 2025-AD010316788). The authors thank the NLP for Biomedical Information Analysis team at the Barcelona Supercomputing Center for their collaboration and valuable discussions. The work was funded by the Île-de-France region through the AI4IDF scholarship.

References

[Agarwal *et al.*, 2022] Dhruv Agarwal, Rico Angell, Nicholas Monath, and Andrew McCallum. Entity Linking via Explicit Mention-Mention Coreference Modeling. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, July 2022.

[Borchert *et al.*, 2024] Florian Borchert, Ignacio Llorca, and Matthieu-P. Schapranow. Improving biomedical entity linking for complex entity mentions with LLM-based text simplification. *Database*, 2024, February 2024.

[Cao *et al.*, 2021] Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. Autoregressive Entity Retrieval. In *International Conference on Learning Representations*, October 2021.

[Chen *et al.*, 2025] Haihua Chen, Ruochi Li, Ana Cleveland, and Junhua Ding. Enhancing data quality in medical concept normalization through large language models. *Journal of Biomedical Informatics*, 165:104812, May 2025.

[Chen *et al.*, 2026] Haihua Chen, Yuhan Zhou, Ruochi Li, Aryan Murthy Illa, Ana Cleveland, and Junhua Ding. A comprehensive survey on medical concept normalization: Datasets, techniques, applications, and future directions. *Journal of Biomedical Informatics*, 178:105005, March 2026.

[French and McInnes, 2023] Evan French and Bridget T. McInnes. An overview of Biomedical Entity Linking throughout the years. *Journal of biomedical informatics*, 137:104252, January 2023.

[Grattafiori *et al.*, 2024] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, et al. The Llama 3 Herd of Models, November 2024. arXiv:2407.21783 [cs].

[Josifoski *et al.*, 2023] Martin Josifoski, Marija Sakota, Maxime Peyrard, and Robert West. Exploiting Asymmetry for Synthetic Training Data Generation: SynthIE and the Case of Information Extraction. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, December 2023.

[Kim *et al.*, 2025] Chanhwi Kim, Hyunjae Kim, Sihyeon Park, Jiwoo Lee, Mujeen Sung, and Jaewoo Kang. Learning from Negative Samples in Biomedical Generative Entity Linking. In *Findings of the Association for Computational Linguistics: ACL 2025*, July 2025.

[Lima-López *et al.*, 2023a] Salvador Lima-López, Eulalia Farré-Maduell, Luis Gasco Sanchez, Jan Rodríguez-Miret, and Martin Krallinger. Overview of SympTEMIST at BioCreative VIII: corpus, guidelines and evaluation of systems for the detection and normalization of symptoms, signs and findings from text. In *Proceedings of the BioCreative VIII Challenge and Workshop: Curation and Evaluation in the Era of Generative Models*, November 2023.

[Lima-López *et al.*, 2023b] Salvador Lima-López, Eulàlia Farré-Maduell, Luis Gasco, Anastasios Nentidis, Anastasia Krithara, Georgios Katsimpras, Georgios Paliouras, and Martin Krallinger. Overview of MedProcNER task on medical procedure detection and entity linking at BioASQ 2023. *Working Notes of CLEF*, 2023.

[Lin *et al.*, 2025] Zhenxi Lin, Ziheng Zhang, Jian Wu, Yefeng Zheng, and Xian Wu. Guiding Large Language Models for Biomedical Entity Linking via Restrictive and Contrastive Decoding. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, November 2025.

[Liu *et al.*, 2021a] Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. Self-Alignment Pretraining for Biomedical Entity Representations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021.

[Liu *et al.*, 2021b] Fangyu Liu, Ivan Vulić, Anna Korhonen, and Nigel Collier. Learning Domain-Specialised Representations for Cross-Lingual Biomedical Entity Linking. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, August 2021.

[Miranda-Escalada *et al.*, 2022] Antonio Miranda-Escalada, Luis Gasco, Salvador Lima-López, Eulàlia Farré-Maduell, Darryl Estrada, Anastasios Nentidis, Anastasia Krithara, Georgios Katsimpras, Georgios Paliouras, and Martin

Krallinger. Overview of DisTEMIST at BioASQ: Automatic detection and normalization of diseases from clinical texts: results, methods, evaluation and multilingual resources. In *Working Notes of Conference and Labs of the Evaluation (CLEF) Forum. CEUR Workshop Proceedings*, 2022.

[Mohan and Li, 2018] Sunil Mohan and Donghui Li. MedMentions: A Large Biomedical Corpus Annotated with UMLS Concepts. In *Automated Knowledge Base Construction (AKBC)*, November 2018.

[Neumann *et al.*, 2019] Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, August 2019.

[Newman-Griffis *et al.*, 2021] Denis Newman-Griffis, Guy Divita, Bart Desmet, Ayah Zirikly, Carolyn P. Rosé, and Eric Fosler-Lussier. Ambiguity in medical concept normalization: An analysis of types and coverage in electronic health record datasets. *Journal of the American Medical Informatics Association: JAMIA*, 28(3):516–532, March 2021.

[Névéol *et al.*, 2014] Aurélie Névéol, Cyril Grouin, Jeremy Leixa, Sophie Rosset, and Pierre Zweigenbaum. The QUAERO French Medical Corpus: A Ressource for Medical Entity Recognition and Normalization. In *Proc of BioTextMining Work*, 2014.

[Remaki *et al.*, 2025] Adam Remaki, Jacques Ung, Pierre Pages, Perceval Wajsburt, Elise Liu, Guillaume Faure, Thomas Petit-Jean, Xavier Tannier, and Christel Gérardin. Improving Phenotyping of Patients With Immune-Mediated Inflammatory Diseases Through Automated Processing of Discharge Summaries: Multicenter Cohort Study. *JMIR Medical Informatics*, 13(1):e68704, April 2025.

[Savova *et al.*, 2010] Guergana K Savova, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, and Christopher G Chute. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513, September 2010.

[Tang *et al.*, 2020] Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. Multilingual Translation with Extensible Multilingual Pretraining and Finetuning, August 2020. arXiv:2008.00401 [cs].

[Vashishth *et al.*, 2021] Shikhar Vashishth, Denis Newman-Griffis, Rishabh Joshi, Ritam Dutt, and Carolyn P. Rosé. Improving broad-coverage medical entity linking with semantic type prediction and large-scale datasets. *Journal of Biomedical Informatics*, 121:103880, September 2021.

[Vollmers *et al.*, 2025] Daniel Vollmers, Hamada Zahera, Diego Moussallem, and Axel-Cyrille Ngonga Ngomo. Contextual Augmentation for Entity Linking using Large Language Models. In *Proceedings of the 31st International Conference on Computational Linguistics*, January 2025.

[Wang *et al.*, 2023] Yi Wang, Corina Dima, and Steffen Staab. Wikimed-DE: Constructing a silver-standard dataset for german biomedical entity linking using wikipedia and wikidata. In *The 4th Wikidata Workshop*, 2023.

[Wu *et al.*, 2020] Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. Scalable Zero-shot Entity Linking with Dense Entity Retrieval. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, November 2020.

[Xiao *et al.*, 2023] Zilin Xiao, Ming Gong, Jie Wu, Xingyao Zhang, Linjun Shou, and Daxin Jiang. Instructed Language Models with Retrievers Are Powerful Entity Linkers. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, December 2023.

[Xin *et al.*, 2025] Amy Xin, Yunjia Qi, Zijun Yao, Fangwei Zhu, Kaisheng Zeng, Bin Xu, Lei Hou, and Juanzi Li. LLMAEL: Large Language Models are Good Context Augmenters for Entity Linking. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, CIKM '25, November 2025.

[Xu *et al.*, 2007] Hua Xu, Peter D Stetson, and Carol Friedman. A study of abbreviations in clinical notes. In *AMIA annual symposium proceedings*, volume 2007, page 821, 2007.

[Ye and Mitchell, 2025] Christophe Ye and Cassie S. Mitchell. LLM as Entity Disambiguator for Biomedical Entity-Linking. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 301–312, Vienna, Austria, July 2025. Association for Computational Linguistics.

[Yuan *et al.*, 2022a] Hongyi Yuan, Zheng Yuan, and Sheng Yu. Generative Biomedical Entity Linking via Knowledge Base-Guided Pre-training and Synonyms-Aware Fine-tuning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, July 2022.

[Yuan *et al.*, 2022b] Zheng Yuan, Zhengyun Zhao, Haixia Sun, Jiao Li, Fei Wang, and Sheng Yu. CODER: Knowledge-infused cross-lingual medical term embedding for term normalization. *Journal of Biomedical Informatics*, 126:103983, February 2022.

[Zhang *et al.*, 2022] Sheng Zhang, Hao Cheng, Shikhar Vashishth, Cliff Wong, Jinfeng Xiao, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Knowledge-Rich Self-Supervision for Biomedical Entity Linking. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, 2022.