SynCABEL : Synthetic Contextualized Augmentation for Biomedical Entity Linking

Anonymous ACL submission

Abstract

Biomedical Entity Linking (BEL) is a fundamental task in biomedical text mining that involves mapping medical mentions in free text to standardized concepts in knowledge bases such as the Unified Medical Language System (UMLS). For instance, BEL enables the secondary use of electronic health records by transforming unstructured clinical notes into structured data, supporting applications such as medical research and clinical decision support. Despite its importance, BEL remains challenging due to the variability and ambiguity of clinical language. Effectively resolving such complexity requires context-aware models, which in turn rely on contextualized training data. However, obtaining such data at scale is difficult given the limited availability of annotated corpora. To address these issues, we introduce **SynCABEL** (Synthetic Contextualized Augmentation for Biomedical Entity Linking), a novel approach that leverages large language models to generate realistic contextual sentences for UMLS concepts, thereby enhancing training for context-aware models. SynCABEL achieves new state-of-the-art results on the QUAERO French Medical Corpus and obtains the highest overall average score across all evaluated datasets, demonstrating its effectiveness and robustness in handling rare and ambiguous concepts. Synthetic datasets, models, and code are released openly to support reproducibility and future research.

1 Introduction

Biomedical Entity Linking (BEL) is a fundamental task in biomedical text mining. It is defined as the process of mapping medical concepts mentioned in free-text clinical notes, research articles, and other biomedical texts to standardized concepts in structured knowledge bases (KBs) or ontologies, such as the Unified Medical Language System (UMLS) (Bodenreider, 2004).

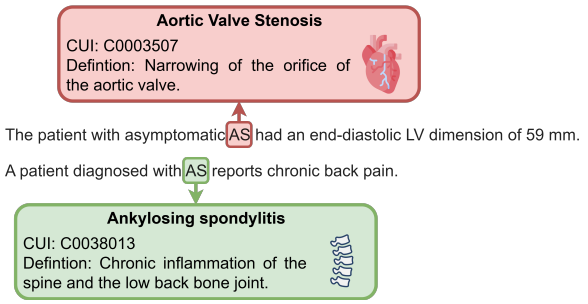


Figure 1: Illustration of the ambiguity problem in Biomedical Entity Linking. The abbreviation “AS" can refer to either “Ankylosing spondylitis" (CUI: C0038013) or “Aortic Valve Stenosis" (CUI: C0003507), requiring contextual cues for resolution. CUI: Concept Unique Identifier

BEL plays a central role in enabling the secondary use of electronic health records by converting unstructured free-text variables into structured concepts. By complementing existing structured data, it supports more comprehensive patient phenotyping and lays the groundwork for downstream applications such as accurate cohort selection, meaningful patient similarity analysis, and more inclusive retrospective studies (Remaki et al., 2025).

Despite its critical role, BEL faces significant challenges. Clinical language exhibits high variability and semantic ambiguity. Concepts often have multiple synonyms (e.g., "myocardial infarction" and "heart attack"), and many terms can be polysemous (e.g., “discharge" can refer to hospital discharge, an administrative act of releasing a patient or pathologic discharge, a fluid emitted from a body orifice). This ambiguity necessitates contextual disambiguation, as illustrated with another example in Figure 1. This challenge is prominent, as Newman-Griffis et al. (2021) demonstrate that exact string matching to UMLS terms leads to over 50% ambiguous matches across several annotated datasets. While context-aware models offer a poten-

tial solution, their training demands annotated data, which is both scarce and expensive to produce due to the need for expert annotation and strict privacy constraints that limit access to real-world clinical data.

To address these limitations, we present **Syn**thetic **C**ontextualized **A**ugmentation for **B**iomedical **E**ntity **L**inking (**SynCABEL**), a novel system that automatically generates realistic contextual sentences for UMLS concepts. By leveraging the generative capabilities of large language models (LLMs), our approach produces contextual synthetic data that enables the effective training of context-aware models. Our key contributions are as follows:

- We release the first open-access, multilingual synthetic dataset for BEL, encompassing all UMLS concepts with definitions available in both English and French.
- We achieves state-of-the-art (SOTA) performance on multiple BEL benchmarks.

2 Related works

BEL can be framed as a massively multiclass classification problem due to the vast size of biomedical terminologies. For instance, the UMLS comprises over 3.5 million concepts and 13.5 million unique concept names spanning around 200 source vocabularies, making the task inherently challenging. Research on BEL has evolved through three major phases: rule-based dictionary matching, context-free neural methods, and contextualized linking models.

2.1 Rule-based dictionary matching

Early biomedical entity linking has relied on exact or approximate string matching against predefined dictionaries. Tools such as KnowledgeMap Concept Identifier (Denny et al., 2003), MetaMap (Aronson and Lang, 2010), cTAKES (Savova et al., 2010), and SciSpacy (Neumann et al., 2019) have been widely integrated into phenotyping pipelines (Liao et al., 2010; Carroll et al., 2011). These systems match surface forms of mentions to candidate concept names using heuristic rules and lexical resources.

However, rule-based methods are inherently limited by their reliance on static dictionaries. They fail to generalize to unseen or ambiguous mentions, and cannot leverage contextual information. It is an important limitation when dealing with homonyms or acronyms commonly found in biomedical texts.

2.2 Context-free neural methods

The introduction of deep contextual embeddings, particularly Transformers (Vaswani et al., 2017), has significantly advanced normalization tasks. SapBERT (Liu et al., 2021) and CODER (Yuan et al., 2022c) are prominent bi-encoder models trained on UMLS synonym pairs using a self-supervised approach, meaning they do not require manually annotated data. Pairs of terms referring to the same concept (e.g., “myocardial infarction" and “heart attack") are treated as positive examples, while terms referring to different concepts serve as negative examples. These models are trained with contrastive learning to pull synonymous mentions closer together and push unrelated ones further apart into a shared embedding space.

While these models are efficient at inference time and perform well on alias matching, they are trained only on isolated synonym pairs, without access to surrounding textual context. As a result, they are less effective at disambiguating mentions in cases involving acronyms or polysemous terms, where contextual cues are essential for accurate resolution.

2.3 Contextualized methods

Unlike alias-based approaches, contextualized BEL methods leverage the mention’s surrounding text to improve disambiguation. Two main families of approaches have emerged:

2.3.1 Bi-encoder models

These models encode both the mention, including its surrounding context, and the candidate entities within a shared embedding space. Although this architecture resembles context-free neural methods, a key distinction is that entity embeddings are learned to capture contextual information during training. MedLinker (Loureiro and Jorge, 2020) is trained on MedMentions (Mohan and Li, 2019), the largest manually annotated biomedical dataset. It represents each entity by averaging the contextualized embeddings of all its associated mentions and relies on approximate dictionary matching for unseen entities. ArboEL (Agarwal et al., 2022) adapts the general-domain bi-encoder BLINK (Wu et al., 2020) to the biomedical domain. It introduces a novel graphical arborescence objective, which models cross-document coreference through directed

spanning trees, achieving new SOTA results on MedMentions. To address the scarcity of French biomedical resources, MLNorm (Wajsbürt et al., 2021) leverages multilingual terminologies and a two-step training strategy that combines distant supervision from UMLS with targeted concept embedding learning. It achieves SOTA performance on the QUAERO French Medical Corpus (Névéol et al., 2014), the largest French BEL dataset.

2.3.2 Autoregressive (generative) models

These treat entity linking as a text generation problem. BioBART (Yuan et al., 2022a), pretrained on PubMed abstracts, achieves strong performance on biomedical generation tasks such as NER and entity linking. GenBioEL (Yuan et al., 2022b) extends this by training a model to generate UMLS concept names using a KB-guided pretraining scheme with synonyms and definitions. It also applies constrained decoding to ensure valid outputs. ANGEL (Kim et al., 2025) goes further by training generative models using hard negative samples, enabling the model to better distinguish entities with similar surface forms. Unlike bi-encoders, generative models avoid encoding all candidate entities, which reduces memory usage. (Cao et al., 2021).

2.4 Data augmentation

Overall, contextualized biomedical entity linking models benefit from leveraging surrounding text but remain limited by the scarcity of high-quality annotated data. This has driven interest in alternative data augmentation strategies.

2.4.1 Weak supervision

Recent work in biomedical entity linking has explored weak supervision to mitigate the scarcity of human-annotated corpora. Zhang et al. (2022) extract mentions with surrounding context from 14 million PubMed articles using exact string matches with UMLS synonyms. Vashishth et al. (2021) extend this line of work by constructing WikiMed and PubMedDS, which automatically link mentions in Wikipedia articles and PubMed abstracts using heuristic patterns and UMLS mappings. While these datasets span millions of mentions and provide useful training signals, they suffer from several limitations: synonym ambiguities can introduce noise, concept coverage is often incomplete, and the extracted contexts are not always representative of real-world clinical text.

2.4.2 Template-based generation

Yuan et al. (2022b) leverage UMLS definitions as textual templates, which are then filled with synonyms to create training instances. This method helps encode entity semantics without relying on labeled data. However, since these examples are generated from templates, they lack context variability.

2.4.3 Synthetic data

More recently, large language models (LLMs) have been used to synthesize or augment entity linking datasets. Outside the biomedical domain, LLMaEL (Xin et al., 2024) employs a generative model to produce detailed descriptions of Wikipedia concepts, which are then injected into context to guide downstream linking. This approach offers control over both the target concept and the type of surrounding context.

3 Our methods

3.1 Problem statement

Let $\mathcal{E}$ represent the set of target concepts from a given knowledge base. Each concept $e \in \mathcal{E}$ is represented by a Concept Unique Identifier CUI_e, a semantic group t_e 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 t_e, our objective is to predict the correct concept e_m that the mention m refers to.

3.2 Generative biomedical entity linking

In this study, we adopt a generative model implemented using a transformer encoder–decoder architecture, following the formulation introduced by Cao et al. (2021). The model takes as input x a sequence defined as:

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

where c_l and c_r denote the left and right context, m is the mention to link (enclosed in square brackets), t_e is the corresponding semantic group (enclosed in braces). The special token SEP separates the sentence context from a template phrase, while the trailing expression "m is" explicitly cues the model to generate the entity corresponding to mention m. The corresponding output sequence y is defined as:

$$y = e_m$$

where e_m is the entity corresponding to mention m. The model is trained to maximize the conditional

likelihood of the output sequence y given the input x:

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

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 CUIs 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.

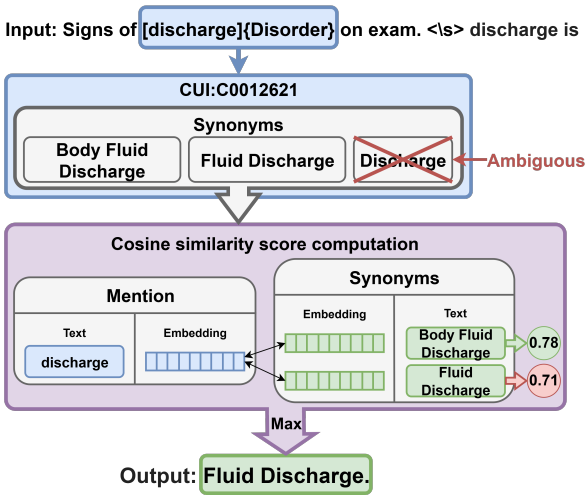


Figure 2: Example of the disambiguation process used to select a natural representation e_m = "Fluid Discharge" for a concept e with CUI_e = "C0012621" and a mention m = "discharge". CUI: Concept Unique Identifier

3.4 Synthetic data generation

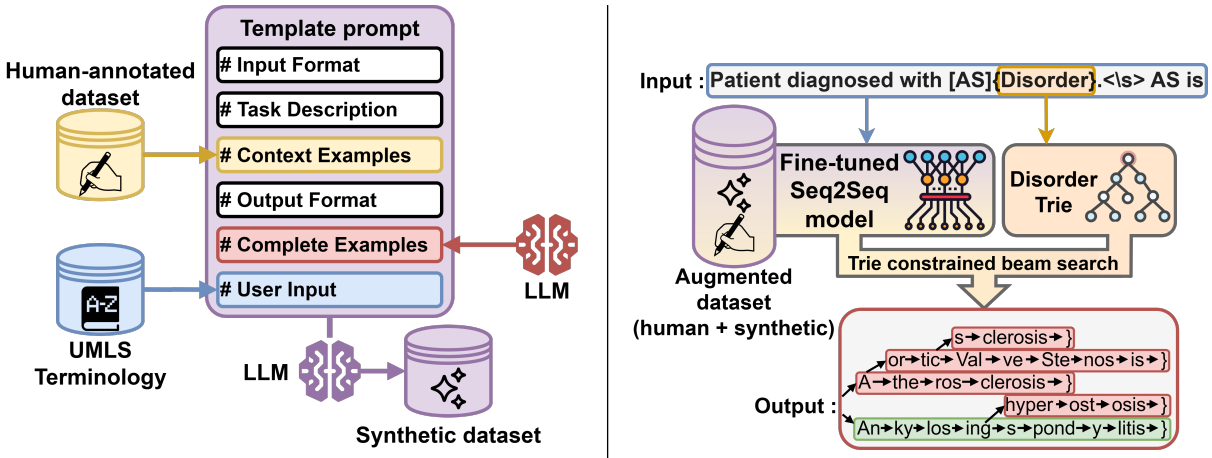
Given that the UMLS encompasses on the order of millions of concepts, no human-annotated dataset comprehensively covers this space. To improve concept coverage while maintaining contextual realism, we generate synthetic training data using a structured prompt-based approach.

For a given concept e, our method generates contextual sentences that capture diverse, natural ways of expressing e. As illustrated in Figure 3(a), this is achieved by constructing a structured prompt composed of three key components: (i) contextual examples from a human-annotated dataset, (ii) Title, semantic group, semantic type, definitions and synonyms of the target concept e from the UMLS terminology, and (iii) complete examples generated by an LLM. The task description section explicitly structures the generation process into two sequential tasks: first, selecting and generating mentions of the concept. Second, generating contextualized sentences for each selected or newly generated mention, aligned with the style of the provided human-annotated dataset examples. This structured two-step prompt is provided to an LLM, which produces the final contextualized sentences. An illustrative example of this prompt design is provided in AppendixA.

3.5 Guided inference

The greedy decoding algorithm (Germann, 2003), which selects the highest-probability token at each step, provides a straightforward approach for sequence generation. However, this method often produces invalid entities that are absent from the UMLS vocabulary.

To address this limitation, we adopted a guided inference mechanism introduced by Cao et al. (2021) that ensures valid concept generation at each inference step. This approach illustrated in Figure 3 (b), operates in two phases. First, the UMLS vocabulary is filtered using the semantic group t_e of the target entity (e.g., *Disorder*). This filtered vocabulary is then structured as a trie (Cormen et al., 2022), where each path represents a disambiguated synonym that can be linked to a single concept such as *Aortic Sclerosis* or *Ankylosing Spondylitis* for the input "AS". 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*).



(a) **Synthetic data generation**.

(b) **Guided inference**.

Figure 3: Overview of SynCABEL. LLM: large language model; UMLS: Unified Medical Language System.

4 Experiments

4.1 Human-annotated datasets

For evaluation, we relied on two human-annotated datasets:

- MedMentions-ST21pv (Mohan and Li, 2019) (MM-ST21pv), a curated subset of the MedMentions corpus, consisting of 4,392 PubMed abstracts in English. Currently, MedMentions is the largest publicly available, human-annotated dataset for BEL.
- The QUAERO French Medical Corpus (Névéol et al., 2014) (QUAERO), the largest French BEL dataset composed of two subsets: QUAERO-MEDLINE, based on MEDLINE titles, and QUAERO-EMEA, derived from European Medicines Agency documents.

	MM-ST21PV	QUAERO-MEDLINE	QUAERO-EMEA
# Docs	4,392	2,498	38
# Mentions	203,282	9,074	7,159
# Concepts	25,419	4,340	1,243
Language	English	French	French
Doc Type	PubMed abstracts	MEDLINE titles	EMEA documents

Table 1: Summary statistics of the human-annotated datasets. MM-ST21pv: MedMentions-ST21pv.

These datasets were chosen for their extensive coverage of UMLS semantic groups and their representation of UMLS vocabularies across two languages. Table 1 summarizes the key characteristics of these datasets. In addition, this evaluation required dedicated UMLS subsets aligned with each dataset's annotation guidelines. For each dataset, we constructed a tailored UMLS subset by applying its specific semantic filters, as detailed in Appendix B.

4.2 Synthetic dataset

We constructed two synthetic datasets to complement the human-annotated corpora. For each dataset, we began by using GPT 4o (OpenAI et al., 2024) to generate two complete examples for few shot prompting. This step was performed only once per dataset. These examples were then incorporated into prompts, with one prompt crafted for each concept. Each of these prompts was given to LLaMA 3.3 70B (Grattafiori et al., 2024), which generated three examples per concept to build the synthetic dataset. We restricted generation to concepts that had at least one available definition. The resulting datasets, SynthMM and SynthQUAERO, were designed to reflect the context of the MM-ST21pv and QUAERO datasets, respectively. Table 2 presents key statistics of the generated datasets.

Dataset	SynthMM	SynthQUAERO
# Docs	460,683	396,636
# Mentions	460,878	396,914
# Concepts	153,374	132,089

Table 2: Summary of synthetic datasets.

4.3 Base models

We evaluated several encoder-decoder architectures as base models, including BART-large (Lewis et al., 2019), BioBART-large (Yuan et al., 2022a) (a biomedical adaptation of BART-large), mT5-large (Xue et al., 2021) and mBART-large (Tang et al., 2020) (both pre-trained on multilingual datasets), and ANGEL (Kim et al., 2025) (pre-trained on UMLS with hard negatives). A detailed performance comparison is provided in Appendix C. As mBART-large demonstrated the best performance across the datasets, it was selected as the base model for all subsequent experiments.

4.4 Adaptive concept representation for training

We evaluated our proposed adaptive concept representation method against a static baseline using UMLS preferred titles. For the adaptive approach, we investigated two distinct representation strategies: the character 3-gram TF-IDF (Neumann et al., 2019) and another using CODER-all embeddings (Yuan et al., 2022c). The results, summarized in Table 3, show that the adaptive method with TF-IDF outperforms both the static baseline and the adaptive method using CODER-all embeddings. Based on this superior performance across all benchmarks, we selected the TF-IDF-based adaptive representation for all subsequent experiments.

	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 3: Recall@1 scores for different concept representation methods using mBART-large. The model is supervised only on the original dataset (i.e., without data augmentation), and inference is performed using the guided inference method.

4.5 Synthetic data generation

Synthetic data generation was performed using a single H100 GPU with 80GB of memory. GPT-4o was run with default parameters: temperature and top-p both set to 1. LLaMA 3.3 70B was run with a temperature of 0.6 and top-p set to 0.9. On average, generation took 4.23 seconds per CUI. For the SynthMM dataset, which contains 153,374 CUIs, the total generation time was approximately 180 hours, or about seven and a half days.

4.6 Training

All models are trained using standard seq2seq optimization with teacher forcing (Williams and Zipser, 1989) and label smoothing (Szegedy et al., 2016). All trainings were performed in parallel on 4 NVIDIA H100 GPUs, each with 80GB of memory. The mBART-large (610.9M parameters) training on the augmented MM-ST21pv combining the original dataset with our synthetically generated examples lasted 13.8 hours for a total of $484,400$ steps. Hyperparameters were selected empirically and are summarized in Appendix D.

4.7 Entity linking inference

BEL inference was performed using a single H100 GPU with 80GB of memory. We applied beam search decoding with the number of beams set to 5, allowing the model to explore multiple candidate sequences and retain the five most probable ones at each step. The temperature and top-p parameters were both empirically set to 1. When using guided decoding with a batch size of 64, the mBART-large model processed the MM-ST21pv test set in 13.1 minutes corresponding to 51.0 mentions per second.

4.8 Main results

Table 4 summarizes the entity linking performance (Recall@1) across three biomedical benchmarks: MM-ST21pv, QUAERO-MEDLINE, and QUAERO-EMEA. The evaluated models are categorized into rule-based (unsupervised), context-free (self-supervised), contextualized bi-encoder (supervised), and contextualized generative (supervised) approaches.

ArboEL (Agarwal et al., 2022) maintains the SOTA on MM-ST21pv with a Recall@1 of 74.2. Our SynCABEL framework which is based on mBART-large, enhanced with guided inference, and trained on an augmented dataset combining the original human-annotated data with synthetic examples sets a new SOTA on the QUAERO corpora and achieves the best overall average score. It obtains Recall@1 scores of 71.5 on MM-ST21pv, 77.1 on QUAERO-MEDLINE, and 75.3 on QUAERO-EMEA.

Model	MM-ST21PV	QUAERO-MEDLINE	QUAERO-EMEA	Average
RULE-BASED (UNSUPERVISED)				
SciSpacy (Neumann et al., 2019)	53.9	41.1	38.0	44.3
CONTEXT-FREE (SELF-SUPERVISED)				
SapBERT (Liu et al., 2021)	51.1	50.6	49.8	50.5
CODER-all (Yuan et al., 2022c)	56.6	58.7	58.1	57.8
CONTEXTUALIZED BI-ENCODER (SUPERVISED)				
ArboEL (Agarwal et al., 2022)	**74.2**	70.9	62.8	69.3
MLNorm* (Wajsbürt et al., 2021)	–	74.0	67.0	–
CONTEXTUALIZED GENERATIVE (SUPERVISED)				
mBART-large (Tang et al., 2020)	65.5	61.5	58.6	61.9
+ Guided inference	70.0	72.8	71.1	<u>71.3</u>
+ SynCABEL	<u>71.5</u> (+1.5)	**77.1** (+4.3)	**75.3** (+4.2)	**74.6**

Table 4: Entity linking performance (Recall@1) on biomedical benchmarks. "*" indicates results reported in the original paper and not reproduced in our experiments. The highlighted row shows our contribution, where the base model is trained on a dataset augmented with synthetic data. 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 three benchmarks.

4.9 ArboEL

To ensure a fair comparison under identical conditions, we reproduced the experiments of Agarwal et al. (2022) using the author's original implementation[1] on our three datasets, using the same UMLS subsets and applying semantic group constraints for candidate generation. We then evaluated the impact of our synthetic data augmentation on this reproduced model. The results, shown in Table 5, indicate that applying the SynCABEL augmentation strategy to ArboEL led to a decrease in performance across all benchmarks. This suggests that the ArboEL architecture, which combines a bi-encoder for candidate generation with a cross-encoder for re-ranking may require a tailored fine-tuning strategy to leverage it effectively. Exploring such a fine-tuning strategy is left for future work. All hyperparameters were kept identical to those reported in the original paper.

Model	MM-ST21PV	QUAERO-MEDLINE	QUAERO-EMEA
ArboEL	74.2	70.9	62.8
+ SynCABEL	67.9 (-6.3)	58.4 (-12.5)	54.2 (-8.6)

Table 5: Performance impact of synthetic data augmentation on ArboEL.

[1] https://github.com/dhdhagar/arboEL

4.10 Scisapcy, SapBERT and CODER-all

For our experiments, SciSpacy[2] was used directly from its official library. The SapBERT[3] and CODER-all[4] models were loaded from their publicly available checkpoints on Hugging Face Hub. To ensure a fair comparison with other contextualized models in our benchmark, we constrained the candidate space for each mention to concepts sharing the same UMLS semantic group.

5 Discussion

5.1 What is the impact of guided inference?

Guided inference substantially improves entity linking accuracy by constraining the model to generate only valid concepts within the input semantic group. This approach yields consistent gains in Recall@1 across all evaluated benchmarks. However, this improvement incurs a computational cost, increasing inference time by approximately 30% compared to unconstrained decoding.

[2] https://allenai.github.io/scispacy/
[3] https://huggingface.co/cambridgeltl/SapBERT-from-PubMedBERT-fulltext
[4] https://huggingface.co/GanjinZero/coder_all

5.2 What is the impact of synthetic augmentation?

SynCABEL leads to substantial performance gains, especially on the QUAERO dataset. As shown in Table 6, on the QUAERO-EMEA subset, performance on unseen mentions improves by 7.4 points, while unseen concepts see an increase of 9.9 points in Recall@1. One possible reason for its stronger impact on QUAERO compared to MedMentions is the smaller size of the dataset, where limited training data makes synthetic augmentation particularly beneficial. Nevertheless, we observed a small performance drop on seen mentions in the QUAERO dataset. While the cause is unclear, one possible explanation is that the small dataset size leads to seen concepts being somewhat diluted by the larger volume of synthetic examples.

Subset	w/o **SynCABEL**	w/ **SynCABEL**	**Sample Size**
MEDMENTIONS-ST21PV			
Overall	70.0	71.5 (**+ 1.5**)	40.14k
Seen Mention	85.6	85.6 (**+ 0.0**)	23.29k (58%)
Seen Concept	79.5	79.5 (**+ 0.0**)	31.57k (79%)
Unseen Mention	48.8	52.6 (**+ 3.8**)	16.85k (42%)
Unseen Concept	37.1	44.2 (**+ 7.1**)	8.57k (21%)
QUAERO-MEDLINE			
Overall	72.8	77.1 (**+ 4.3**)	3.09k
Seen Mention	91.5	91.1 (**- 0.4**)	1.16k (38%)
Seen Concept	89.4	89.7 (**+ 0.3**)	1.72k (56%)
Unseen Mention	63.8	70.3 (**+ 6.5**)	1.93k (62%)
Unseen Concept	56.3	64.5 (**+ 8.2**)	1.37k (44%)
QUAERO-EMEA			
Overall	71.1	75.3 (**+ 4.2**)	2.23k
Seen Mention	94.7	93.6 (**- 1.1**)	1.01k (45%)
Seen Concept	92.9	90.8 (**- 2.1**)	1.19k (53%)
Unseen Mention	56.8	64.2 (**+ 7.4**)	1.22k (55%)
Unseen Concept	51.3	61.2 (**+ 9.9**)	1.04k (47%)

Table 6: Recall@1 performance on three biomedical benchmarks comparing two variants of mBART-large: one trained only on the corresponding human-annotated dataset with guided inference, and one further enhanced with SynCABEL, which is additionally trained on the corresponding synthetic dataset. Results are reported overall and for subsets defined by seen/unseen mentions and concepts. Sample sizes and subset proportions are shown. Differences between the two models are shown in parentheses and highlighted in bold.

5.3 How do generative models compare to bi-encoder architectures in practice?

Our experiments reveal complementary strengths and limitations between encoder architectures and generative models for BEL. The ArboEL model, which combines a bi-encoder with a cross-encoder reranker, performs best on the large, English-language MM-ST21pv dataset. This suggests that the approach is well suited for settings with abundant annotated data. However, its fine-tuning process does not scale efficiently, as performance decrease when trained with more data. Moreover, the method requires storing embeddings for all concepts, which leads to substantial memory demands and preprocessing overhead. Inference is also comparatively slower, taking around 20 minutes on MM-ST21pv due to nearest-neighbor search computations.

In contrast, the generative mBART-based model requires no precomputed embeddings, making it both lighter in memory and faster at inference, completing the same task in 13 minutes even with guided decoding. Generative models also scale more effectively with data augmentation, showing consistent gains on unseen concepts without significant degradation on seen ones. These characteristics make auto-regressive generative models particularly attractive for multilingual or low-resource biomedical settings.

6 Conclusion

In this paper, we introduced SynCABEL, a framework for biomedical entity linking that leverages synthetic data generation to address the critical challenge of limited annotated training data. By harnessing the capabilities of LLMs, SynCABEL generates realistic, contextually rich training examples for UMLS concepts, enabling more effective training of context-aware BEL models. Our experiments show that SynCABEL achieves SOTA performance on the QUAERO French Medical Corpus, with substantial gains on unseen concepts. Notably, its impact is strongest when applied to smaller datasets, highlighting its potential for low-resource languages and domains where annotated biomedical data remain limited. All generated synthetic datasets are released publicly to foster reproducibility and future research.

7 Limitations

While SynCABEL demonstrates strong performance, our work has several limitations that present opportunities for future research.

The first limitation concerns the incomplete coverage of our synthetic data. To ensure quality, we only generated data for UMLS concepts with available definitions, leaving many concepts without synthetic examples. For example, in the MM-ST21pv dataset, 1,625 concepts were still unseen

during training, even after augmentation. This represents a clear improvement compared to the 3,240 unseen concepts in the human-annotated data, but it also highlights a remaining limitation. Therefore, future work could focus on extending the coverage of the synthetic datasets to all concepts.

Second, the evaluation of our synthetic data has so far been purely extrinsic, focusing on its downstream impact on BEL performance. Future work may include an intrinsic assessment conducted by medical experts. Such an evaluation would directly measure the coherence of the generated text, offering deeper insights into the strengths and limitations of the data generation process.

A third limitation is the restriction of our base model to encoder–decoder architectures. In principle, the same guided inference and data augmentation strategy could be applied to modern decoder-only architectures, which now play a central role in biomedical NLP (Chen et al., 2025). Evaluating SynCABEL with such models would therefore be a promising direction for future work.

Finally, the practical deployment of our model in a clinical setting, such as on raw clinical notes, would require a named entity recognition system to detect entity mentions and classify their semantic groups. Additionally, the synthetic data generation would need to use examples from clinical texts rather than abstracts to capture their telegraphic style and heavy use of abbreviations.

8 Ethical considerations

This work relies exclusively on synthetic data, mitigating privacy risks associated with the use of real patient records. No identifiable clinical data were used in any stage of training or evaluation. By avoiding real patient information, our approach ensures that privacy concerns are minimized.

Generating large-scale synthetic data with language models, however, has a notable environmental cost due to GPU energy consumption. We encourage future research to carefully evaluate the trade-off between data volume, model performance, and environmental impact.

All existing artifacts were used in accordance with their licenses and access conditions. QUAERO was used under GFDL for research purposes, MedMentions under CC0, and the UMLS Metathesaurus under its License Agreement, which permits research use with acknowledgment of the National Library of Medicine. Our synthetic datasets were derived solely for research in biomedical entity linking, and their intended use is clearly specified for research purposes. They are released under an Apache 2.0 license, ensuring compatibility with the original access conditions and enabling other researchers to build upon our work.

The data we collected, used, or released should not contain any personally identifying information or harmful content. The synthetic data was generated by an LLM that has been fine-tuned to minimize the risk of producing identifying, harmful, or offensive content. Nevertheless, as with any generative model, unintended outputs cannot be completely ruled out.

The authors used GitHub Copilot for code completion. All code was reviewed and is the authors' responsibility.

References

Dhruv Agarwal, Rico Angell, Nicholas Monath, and Andrew McCallum. 2022. 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*, pages 4644–4658, Seattle, United States. Association for Computational Linguistics.

Alan R Aronson and François-Michel Lang. 2010. An overview of MetaMap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236.

Olivier Bodenreider. 2004. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32(Database issue):D267–270.

Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. 2021. Autoregressive Entity Retrieval. *arXiv preprint*. ArXiv:2010.00904 [cs].

Robert J. Carroll, Anne E. Eyler, and Joshua C. Denny. 2011. Naïve Electronic Health Record Phenotype Identification for Rheumatoid Arthritis. *AMIA Annual Symposium Proceedings*, 2011:189–196.

Qingyu Chen, Yan Hu, Xueqing Peng, Qianqian Xie, Qiao Jin, Aidan Gilson, Maxwell B Singer, Xuguang Ai, Po-Ting Lai, Zhizheng Wang, and 1 others. 2025. Benchmarking large language models for biomedical natural language processing applications and recommendations. *Nature communications*, 16(1):3280.

Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein. 2022. *Introduction to Algorithms, fourth edition*. MIT Press. Google-Books-ID: RSMuEAAAQBAJ.

Joshua C. Denny, Plomarz R. Irani, Firas H. Wehbe, Jeffrey D. Smithers, and Anderson Spickard. 2003. The KnowledgeMap Project: Development of a Concept-Based Medical School Curriculum Database. *AMIA Annual Symposium Proceedings*, 2003:195–199.

Ulrich Germann. 2003. Greedy Decoding for Statistical Machine Translation in Almost Linear Time. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 72–79.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and Yang. 2024. The Llama 3 Herd of Models. *arXiv preprint*. ArXiv:2407.21783 [cs].

Chanhwi Kim, Hyunjae Kim, Sihyeon Park, Jiwoo Lee, Mujeen Sung, and Jaewoo Kang. 2025. Learning from Negative Samples in Biomedical Generative Entity Linking. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 10714–10730, Vienna, Austria. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. *arXiv preprint*. ArXiv:1910.13461 [cs].

Katherine P. Liao, Tianxi Cai, Vivian Gainer, Sergey Goryachev, Qing Zeng-treitler, Soumya Raychaudhuri, Peter Szolovits, Susanne Churchill, Shawn Murphy, Isaac Kohane, Elizabeth W. Karlson, and Robert M. Plenge. 2010. Electronic medical records for discovery research in rheumatoid arthritis. *Arthritis Care & Research*, 62(8):1120–1127. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/acr.20184.

Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. 2021. Self-Alignment Pretraining for Biomedical Entity Representations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4228–4238, Online. Association for Computational Linguistics.

Daniel Loureiro and Alípio Mário Jorge. 2020. MedLinker: Medical Entity Linking with Neural Representations and Dictionary Matching. In *Advances in Information Retrieval*, pages 230–237, Cham. Springer International Publishing.

Sunil Mohan and Donghui Li. 2019. MedMentions: A Large Biomedical Corpus Annotated with UMLS Concepts. *arXiv preprint*. ArXiv:1902.09476 [cs].

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 319–327, Florence, Italy. Association for Computational Linguistics.

Denis Newman-Griffis, Guy Divita, Bart Desmet, Ayah Zirikly, Carolyn P. Rosé, and Eric Fosler-Lussier. 2021. 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.

Aurélie Névéol, Cyril Grouin, Jeremy Leixa, Sophie Rosset, and Pierre Zweigenbaum. 2014. The QUAERO French Medical Corpus: A Ressource for Medical Entity Recognition and Normalization. In *Proc of BioTextMining Work*, pages 24–30.

OpenAI, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, A. J. Ostrow, Akila Welihinda, and Hayes. 2024. GPT-4o System Card. *arXiv preprint*. ArXiv:2410.21276 [cs].

Adam Remaki, Jacques Ung, Pierre Pages, Perceval Wajsburt, Elise Liu, Guillaume Faure, Thomas Petit-Jean, Xavier Tannier, and Christel Gérardin. 2025. Improving Phenotyping of Patients With Immune-Mediated Inflammatory Diseases Through Automated Processing of Discharge Summaries: Multicenter Cohort Study. *JMIR Medical Informatics*, 13(1):e68704. Company: JMIR Medical Informatics Distributor: JMIR Medical Informatics Institution: JMIR Medical Informatics Label: JMIR Medical Informatics Publisher: JMIR Publications Inc., Toronto, Canada.

Guergana K Savova, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, and Christopher G Chute. 2010. 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.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the Inception Architecture for Computer Vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826.

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

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International*

Conference on Neural Information Processing Systems, NIPS'17, pages 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Perceval Wajsbürt, Arnaud Sarfati, and Xavier Tannier. 2021. Medical concept normalization in French using multilingual terminologies and contextual embeddings. *Journal of Biomedical Informatics*, 114:103684.

Ronald J. Williams and David Zipser. 1989. A Learning Algorithm for Continually Running Fully Recurrent Neural Networks. *Neural Computation*, 1(2):270–280.

Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. 2020. Scalable Zero-shot Entity Linking with Dense Entity Retrieval. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6397–6407, Online. Association for Computational Linguistics.

Amy Xin, Yunjia Qi, Zijun Yao, Fangwei Zhu, Kaisheng Zeng, Xu Bin, Lei Hou, and Juanzi Li. 2024. LLMAEL: Large Language Models are Good Context Augmenters for Entity Linking. *arXiv preprint*. ArXiv:2407.04020 [cs].

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint*. ArXiv:2010.11934 [cs].

Hongyi Yuan, Zheng Yuan, Ruyi Gan, Jiaxing Zhang, Yutao Xie, and Sheng Yu. 2022a. BioBART: Pretraining and Evaluation of A Biomedical Generative Language Model. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 97–109, Dublin, Ireland. Association for Computational Linguistics.

Hongyi Yuan, Zheng Yuan, and Sheng Yu. 2022b. Generative Biomedical Entity Linking via Knowledge Base-Guided Pre-training and Synonyms-Aware Fine-tuning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4038–4048, Seattle, United States. Association for Computational Linguistics.

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

Sheng Zhang, Hao Cheng, Shikhar Vashishth, Cliff Wong, Jinfeng Xiao, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2022. Knowledge-Rich Self-Supervision for Biomedical Entity Linking. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 868–880, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

A Example prompt

 # Task Description
Perform these tasks:
1. **Mention Selection & Generation:**
  - First evaluate which existing mentions sound most natural in professional biomedical writing (clinical or research contexts)
  - Keep only the most fluent mentions from the existing list
  - Then generate additional high-quality mentions that:
    * Add natural variations researchers would use
    * Cover different naming conventions (acronyms, full names, functional names)
    * Complement the existing mentions
  - The total should be 3 mentions (mix of kept existing and newly generated)
  - **Avoid**:
    - Commas (`,`), parentheses.
    - ALL_CAPS or underscored terms (e.g., `ALPHA_TOCOPHEROL`).
    - Chemical names.
    - Non-English terms or foreign language mentions.
2. **Contextual Usage Examples:**
  For each of the final 3 mentions:
  - Write one clear sentence (15-30 words) with the mention in brackets in the style of a biomedical abstract.
  - Each must:
    - Include the concept in `[brackets]`.
    - Start with `example: `.
    - Be **distinct in context** (e.g., different study types, organisms, biological systems, or outcomes)
    - Mimic the tone and vocabulary typical of **PubMed biomedical research** papers

 # Input Format
- **CUI**: A unique concept ID
- **Title**: The preferred name or title of the concept
- **Semantic group**: The UMLS semantic group
- **Semantic Type**: The specific semantic type
- **Definitions**: One or more definitions of the concept
- **Mentions**: A list of mentions (may include technical or unnatural terms)

 # Output Format
Final Mention Set
1. Mention 1
2. Mention 2
3. Mention 3

Usage Examples
example: Sentence with [mention 1] in context...
example: Sentence with [mention 2] in context...
example: Sentence with [mention 3] in context...

 # Context Examples
- **<sample 1 from MedMentions-ST21pv>**
- **<sample 2 from MedMentions-ST21pv>**
- **<sample 3 from MedMentions-ST21pv>**

 # Complete Examples
Example 1
User input
- **CUI**: C3539909
- **Title**: Allergic disposition
- **Semantic group**: Disorders
- **Semantic Type**: Disease or Syndrome
- **Definitions**: * 1. The disposition to develop an allergic reaction.
- **Mentions**: 'Allergic disposition', 'Allergic disposition (disorder)'

Final Mention Set
1. Allergic disposition
2. Allergic predisposition
3. Allergy susceptibility

Usage Examples
example: In a prospective birth cohort, maternal vitamin D status and neonatal epigenetic signatures were examined as predictors of [Allergic disposition] at school age.
(...)

Example 2
(...)

 # User Input
- **CUI**: **<Target CUI>**
- **Title**: **<Target Title>**
- **Semantic group**: **<Target Semantic group>**
- **Semantic Type**: **<Target Semantic type>**
- **Definitions**: **<Target Definitions>**
- **Mentions**: **<Target Mentions>**

 # Output
example:

Example of the structured prompt used for synthetic data generation.

B UMLS terminology

We created two distinct subsets of the UMLS. Each subset corresponded to a human-annotated dataset and was consistent with its annotation guidelines.

- For MM-ST21pv, we use the UMLS 2017AA release, further filtered using the 21 Semantic Types and Preferred Vocabularies (ST21-pv) that includes only concepts that belong to one

Model	MM-ST21pv	QUAERO-MEDLINE	QUAERO-EMEA
mT5-large (Xue et al., 2021)	64.1	62.5	64.2
+ Guided inference	68.5	73.2	69.7
+ SynCABEL	69.5 (+ 1.0)	75.6 (+ 2.4)	73.7 (+ 4.0)
ANGEL (Kim et al., 2025)	64.7	60.0	63.4
+ Guided inference	67.2	68.6	67.7
+ SynCABEL	69.4 (+ 2.2)	75.7 (+ 7.1)	74.8 (+ 7.1)
BioBART-large (Yuan et al., 2022a)	66.1	60.6	62.8
+ Guided inference	70.2	69.3	68.3
+ SynCABEL	67.4 (- 2.8)	75.9 (+ 6.6)	75.3 (+ 7.0)
BART-large (Lewis et al., 2019)	61.5	59.4	61.0
+ Guided inference	69.0	70.9	70.0
+ SynCABEL	70.0 (+ 1.0)	74.3 (+ 3.4)	**76.8** (+ 6.8)
mBART-large (Tang et al., 2020)	65.5	61.5	58.6
+ Guided inference	70.0	72.8	71.1
+ SynCABEL	**71.5** (+ 1.5)	**77.1** (+ 4.3)	75.3 (+ 4.2)

Comparison of different base models used in our methods. The table reports entity linking performance (Recall@1) on three biomedical benchmarks: MM-ST21pv, QUAERO-MEDLINE, and QUAERO-EMEA. The highlighted rows show our contributions, where the base models are trained on the dataset augmented with synthetic data. The best results are shown in **bold**, and the second-best results are underlined.

of 21 commonly used biomedical semantic types (or their descendants in the UMLS semantic hierarchy). Additionally, synonyms are restricted to those originating from 18 selected source ontologies.

- For QUAERO, we use the UMLS 2014AA release, filtered to retain only concepts that fall within 10 selected semantic groups.

In both cases, we restrict synonyms to those available in both English and French. Table below summarizes the size of the filtered UMLS subsets used for each dataset, in terms of the number of concepts and synonyms.

	MM-ST21pv	QUAERO
UMLS version	2017AA	2014AA
# Concepts	2,374,392	2,960,365
# Synonyms	4,905,087	6,970,593
# Semantic groups	14	10

Summary of UMLS subsets used.

C Evaluation of base models

We evaluated several base models to identify the best architecture for SynCABEL, with results summarized in the Table above. Among all models tested, the multilingual mBART-large achieved the best scores on MM-ST21pv and QUAERO-MEDLINE and the second-best on QUAERO-EMEA. It was closely followed by BART-large as the second-best performer, demonstrating that general-purpose architectures adapt remarkably well. Conversely, the domain-specific BioBART-large exhibited a small performance decrease on the MedMentions benchmark. This may indicate that the benefits of domain-specific pre-training do not automatically translate

D Training hyperparameters

Epochs	100
Warmup steps	500
Learning Rate	3e-5
Learning Scheduler	Linear
Batch Size	32
Optimizer	AdamW
Adam ϵ	1e-8
Adam β	(0.9, 0.999)
Attention Dropout	0.1
Label Smoothing	0.1

Training hyperparameters used for fine-tuning our BEL models.